

A COMPARATIVE STUDY OF NONPARAMETRIC METHODS FOR PATTERN RECOGNITION

PRICES SUBJECT TO CHANGE

(NASA-CR-130824)	A COMPARATIVE STUDY OF	N73-18203
NONPARAMETRIC METHODS FOR PATTERN		
RECOGNITION Interim Technical Report		
(South Dakota State Univ.) 99 p HC		Unclas
	CSCL 06D G3/08	64320

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
US Department of Commerce
Springfield, VA. 22151

Remote Sensing Institute
South Dakota State University
Brookings, South Dakota

November 1972

Interim Technical Report

A COMPARATIVE STUDY OF NONPARAMETRIC
METHODS FOR PATTERN RECOGNITION

by

Sangkoo F. Hahn

and

Gerald D. Nelson

Associate Professor
Electrical Engineering
and
Staff Specialist
Remote Sensing Institute

to

National Aeronautics and Space Administration
Washington, DC
Grant Number NGL 42-003-007

Remote Sensing Institute
South Dakota State University
Brookings, South Dakota

November 1972

ABSTRACT

The applied research discussed in this report determines and compares the correct classification percentage of the non-parametric sign test, Wilcoxon's signed rank test, and K-class classifier with the performance of the Bayes classifier. The performance is determined for data which have Gaussian, Laplacian and Rayleigh probability density functions. The correct classification percentage is shown graphically for differences in modes and/or means of the probability density functions for four, eight and sixteen samples. The K-class classifier performed very well with respect to the other classifiers used. Since the K-class classifier is a nonparametric technique, it usually performed better than the Bayes classifier which assumes the data to be Gaussian even though it may not be. The K-class classifier has the advantage over the Bayes in that it works well with non-Gaussian data without having to determine the probability density function of the data. However, it should be noted that the data in this experiment was always unimodal.

ACKNOWLEDGEMENTS

The authors wish to thank Mr. Victor I. Myers, Director of the Remote Sensing Institute and his staff for enabling the research to be done. This work was partially supported by NASA Grant NGL 42-003-007, and in cooperation with the Electrical Engineering Department at the South Dakota State University.

TABLE OF CONTENTS

CHAPTER	PAGE
I. INTRODUCTION	1
A. Introduction to the Problem	1
B. Objectives and Significance of the Thesis	2
C. Literature Review	4
D. Organization of the Thesis	5
II. PATTERN RECOGNITION ALGORITHMS	7
A. Parametric Methods	7
1. Bayes' decision rule	8
2. Learning with a teacher	11
3. Learning without a teacher	11
4. When functional forms are not known	12
5. Sequential decision methods	13
6. Summary of the parametric methods	14
B. Nonparametric Methods	17
1. Asymptotic relative efficiency	18
2. One-input tests	19
a. Sign test	19
b. Wilcoxon's signed-rank test	21
c. Sequential nonparametric method	23
3. Two-input tests	24
a. Sign test	25
b. Wilcoxon's signed-rank test	25
c. Rank sum test	26

4. Two-input tests(correlation method)	27
a. Rank correlation method	28
b. Polarity coincidence correlation	28
5. Briefs on nonparametric methods	29
III. BASIC PERFORMANCE COMPARISONS	30
A. Generation of Random Signals and Noises	30
1. Random signals	30
2. Random noises	31
B. Gaussian Distribution Case	34
C. Two-sided Exponential Distribution Case	43
D. Rayleigh Distribution Case	53
E. Complexity of Calculation of Each Algorithm	62
F. Summary of the Chapter	68
IV. APPLICATIONS TO THE MULTI-CLASS PROBLEMS	70
A. Univariate, Multi-class Problems	70
1. Sign test	71
2. Signed-rank test	71
3. Rank sum test	72
B. Multivariate, Two-class Problems	73
C. Multivariate, Multi-class Problems	74
D. Summary of the Chapter	74
V. CONCLUSIONS	76
A. Summary	76
B. Suggestions for Further Study	78

GLOSSARY OF TERMS	79
APPENDIX	81
A. Reduction of a Quadratic Form to a Linear Form	81
B. Estimation of Parameters Using Repetitive Calculations . .	83
BIBLIOGRAPHY	85

LIST OF FIGURES

3-1.	Rayleigh distribution	33
3-2.	Error probabilities of different algorithms with Gaussian distribution (16 features)	38
3-3.	The error probability ratios between Bayes' and nonparametric methods (1.0 for Bayes')	40
3-4.	Theoretical and experimental error probabilities of Gaussian distribution (16 features)	42
3-5.a.	Error probabilities of different algorithms with Laplacian distribution (4 features)	46
b.	Error probabilities of different algorithms with Laplacian distribution (8 features)	47
c.	Error probabilities of different algorithms with Laplacian distribution (16 features)	48
3-6.	Relative changes of error probabilities in accordance with sample sizes	52
3-7.a.	Error probabilities of different algorithms with Rayleigh distribution (4 features)	56
b.	Error probabilities of different algorithms with Rayleigh distribution (8 features)	57
c.	Error probabilities of different algorithms with Rayleigh distribution (16 features)	58
3-8.	Two-input nonparametric tests of Rayleigh distribution . .	59
3-9.	Relative changes in error probabilities for different mode values	61

3-10.a.	Changes of α - and β -error probabilities according to thresholds in signed-rank test (Difference in mode values = 0.3. Rayleigh distribution)	63
b.	Changes of α - and β -error probabilities according to thresholds in signed-rank test (Difference in mode values = 0.1. Rayleigh distribution)	64
c.	Changes of α - and β -error probabilities according to thresholds in sign test (Difference in mode values = 0.3. Rayleigh distribution)	65
d.	Changes of α - and β -error probabilities according to thresholds in sign test (Difference in mode value = 0.1. Rayleigh distribution)	66

LIST OF TABLES

II-1.	Available information and possible algorithms of the four cases	16
III-1.	Thresholds for each test	37
III-2.	Predetermined and experimental α -error	41
III-3.	Threshold for each sample size and mean difference	44
III-4.	Average error probabilities for each algorithm with different sample sizes	45
III-5.	Predetermined and experimental probabilities of error in exponential distribution case	51
III-6.	Calculations involved in each algorithm	67

CHAPTER I

INTRODUCTION

A. Introduction to the Problem

Pattern recognition theory has a wide range of applications in radar [32], sonar [5], imagery recognition, and alpha-numeric character identification [8]. The recognition of patterns is accomplished by many different algorithms. They process the input data according to the algorithms in order to draw conclusions. Since a pattern recognition problem is usually concerned with classifying a set of input data into one of many classes, the resultant conclusion is the designation of the input data set to a certain class. In order to use the algorithms they are usually implemented as software or hardware. Therefore, large quantities of data can be processed rapidly and complex data inputs are reduced to outputs which provide a simple, understandable result to a user of the algorithm.

There seems to be no limit for the development of the algorithms. Any algorithm which is useful for the solution to a recognition problem can be included in the field of pattern recognition algorithms. Many algorithms are mathematically well defined and developed to meet the conditions of the problems. There are, however, two general approaches to the recognition problems [38]. One is to treat the problem in deterministic sense, while the other starts with the statistical point of view. The statistical approaches can again be subdivided into parametric and nonparametric methods.

Nonparametric methods, which have attracted the attention of many

investigators recently, have advantages over parametric methods in their relative insensitivity to the changes of the input statistics and no need for a priori information about the parameters of the probability density function (pdf) [2].

One of the problems encountered at the Remote Sensing Institute of South Dakota State University is the recognition of crops on the film, which is exposed at various altitudes. Since the photographic imagery is affected by complex set of factors, the nonparametric methods seem to be appealing to investigators in search for an appropriate recognition algorithm.

Nonparametric methods have inherent drawbacks and it is necessary to compare these methods to those of parametric methods which have already been proposed and used. Many works on nonparametric area have appeared as indicated by the literature review presented in this report. Many of the authors show the good aspects of these nonparametric methods with relatively little about the limitations of their use, especially in the practical situations. A study on the comparative performance of nonparametric methods with respect to the parametric methods is necessary.

B. Objectives and Significance of the Thesis

The main objective of this thesis is to determine and compare the classification error probabilities of several nonparametric methods to parametric ones in practical or near practical conditions using computer simulation. The usefulness of the Asymptotic Relative Efficiency (ARE) is also observed. The ARE is used to compare one algorithm with another in the limit case conditions which are far from practical circumstances. Details of the ARE concept is given in Chapter II and

some of the literatures [7], [22].

The next objective is the investigation of the complexity of the several algorithms studied. Many investigators are implementing their algorithms on computers, and since the computer time is determined by the complexity of the algorithm, a very crucial aspect of any algorithm is its complexity. If the data are processed by other than computer, the hardware of the system required will become more expensive and complicated as the calculation gets more complex. In this respect, the calculation problem is studied.

The previously stated objectives are performed extensively with two-class problems, but the actual classification problem in imagery recognition usually is a multi-class one. Hence, the generalization of the two-class problem to the multi-class one is studied as a minor objective.

One of the important aspects of this work is that the performance of each algorithm with various data distribution conditions can be found in very practical, not theoretical, circumstances. The adoption of a method as a data processing algorithm by a designer of the system can be based more positively on the results of this work. The merits and the limitations of the nonparametric methods are also determined by the actual handling of data through each method.

The effects of sample sizes and signal-to-noise ratios on error probabilities are experimented to give more insight into the algorithm and to see various situational behavior of the method. Through the experiments, determining a nonparametric threshold happens to be an important matter in actual applications of algorithms. This is also

studied and a specific result is drawn.

C. Literature Review

Bradley [1] talks about the justification of using nonparametric methods in many cases. His book is also an excellent source of general information on the nonparametric methods. Several useful cases of nonparametric tests are treated in the works of Carlyle and Thomas [2] and Thomas [2].

Mathematical aspects of nonparametric methods are handled by Fraser [7]. Kraft and van Eeden [17] approach the nonparametric method in a unique fashion using treatment and effect concept. Over 3000 nonparametric references are listed in the work of Savage [21]. A determination of probability density function of sequential rank vector is done by Fu [8], and Fu and Chien [9]. More work on the sequential nonparametric method is given by Chadwick and Kurz [3].

The detailed process of determining the ARE of some nonparametric algorithms with respect to the Student's t-test is given in the famous work of Hodges and Lehmann [15]. They showed that the ARE of the nonparametric rank sum method compared to the t-test never falls below 0.864.

Feustel and Davisson [5] report that mixed statistics is a good way of compromising between calculation complexity and performance efficiency. Daly and Rushforth [4] compare the ARE of nonparametric to parametric optimal detector in the Gaussian and non-Gaussian distribution. It was shown that nonparametric methods are more flexible than the corresponding optimal detectors in ARE sense.

Fralick and Scott [6] deal with the nonparametric nearest-neighbor

method to estimate the Bayes' risk. It is proven by Groeneveld [10] that the method based on the correlation of the signs of differences of observed data has an efficiency exceeding more than unity compared to the parametric method under certain noise distributions.

A procedure is reported by Kanefsky and Thomas [16] that modifies given sampled-data parametric detectors to asymptotically nonparametric ones. Applications of the K-S test to a signal detection problem are performed by Millard and Kurz in their two similar works [18], [19].

D. Organization of the Thesis

Since the nonparametric methods are compared to the parametric Bayes' classifier, a brief review of parametric and nonparametric methods is provided in Chapter II.

Part A of Chapter II deals with parametric methods according to the available a priori knowledge of the probability density function. Part B of the chapter starts with the definition of ARE and explains one-input nonparametric methods as well as the two-input case. Correlation methods are also discussed.

In Chapter III the two-class problem is used to test the performance of nonparametric and Bayes' classifiers. A fixed sample size of 16 is used for each of the five different conditions of the separations of means for the Gaussian data. For the double-sided exponential and Rayleigh distribution cases there are three different sample sizes used for each of the different signal level separations. The different error probabilities for different nonparametric thresholds are also experimented to see the effect of threshold values.

The multi-class problems are treated in Chapter IV. The univariate

multi-class, the multivariate two-class and the multivariate multi-class problem are considered separately in that order.

The conclusions of the thesis work and the suggestions for further research are discussed in the last chapter.

CHAPTER II

PATTERN RECOGNITION ALGORITHMS

A. Parametric Methods

In general, the first decision that should be made by a designer of a system to solve a pattern recognition problem is to make a choice of an algorithm. The designer can choose between a deterministic and a statistical algorithm. A deterministic procedure which has been also very important and well developed [35] will not be discussed here except for the relationships with the statistical one. The statistical approach can be conveniently subdivided into parametric and nonparametric algorithms. A parametric method makes use of the parameters of the probability density function (pdf) or the distribution of input data. The distribution information may not be complete and it is necessary to estimate the parameters. Reasonable assumptions and convenient derivations can be made quite often to make the problem of parameter estimation easier. The question of how good the approximation is compared to the original is not simple to answer. The nonparametric statistical methods will be discussed in part B of this chapter.

The parametric methods can be studied in several cases according to the type or combinations of types of available information [12]. The first type of data information gives only the form of the distribution but not the parameters, θ . In the second type of information, the parameter values are also given in addition to the functional form of the distribution, hence, complete information is furnished. In the third type of data information, neither the functional form of the distribution nor the parameters are given but only a set of samples

from known classes is provided. In this third type, the samples should be utilized to estimate the distribution. The last type of information gives only the samples without any a priori information. This fourth type is the most difficult and probably the most general situation in which pattern recognition algorithms have to be developed. The data samples are used to determine possible decision boundaries. New input data can be classified as soon as the decision boundaries are determined.

While these parametric methods are straightforward and mathematically eligible for deeper analyses, they also have many shortcomings. In many instances, little or almost no prior information about the input data is given. It will be very tedious and time consuming to evaluate the distribution. Even if it is possible to spare the time and labor to figure out the distribution, it may not be easy to represent the distribution with a finite number of parameters because of the complexity of the distribution shape. In the following sections, each case in connection with the data types is studied further.

1. Bayes' decision rule

Consider the case where the distribution is completely known and there are only two classes to classify from. This is the case where the combined information of the data type one and two is furnished. Let the conditional probability density function of class 0 and class 1 be $f(\underline{x}/H_0)$ and $f(\underline{x}/H_1)$, respectively. \underline{x} is the given set of data represented in vector form with n elements and H_0 is the null hypothesis that the data set is from class 0 instead of the alternative H_1 that the data set is from class 1. The most widely accepted decision

criterion is the maximum likelihood ratio. The ratio of the two distributions is compared to a certain threshold of value C . If the ratio exceeds the threshold the hypothesis H_0 is accepted, otherwise alternative H_1 is accepted. It can be written as follows:

$$L(\underline{x}) = f(\underline{x}/H_0) / f(\underline{x}/H_1)$$

and if

$$L(\underline{x}) \geq C \rightarrow H_0 \text{ is accepted or}$$

if $L(\underline{x}) < C \rightarrow H_1$ is accepted.

To determine the bias C is the responsibility of the investigator. The Bayes' decision rule determines the threshold by the a priori probability of class i , $p(i)$, and the cost of making decisions of the class, K_i , as

$$C = p(0)K_0 / p(1)K_1$$

where $p(0)$, $p(1)$ and K_0 , K_1 are assumed known.

The Bayes' decision optimally minimizes the overall risk of making errors. The fundamental Neymann-Pearson criterion requires β to be a minimum for a fixed value of α . It is shown that the likelihood ratio test given above will satisfy the Neymann-Pearson criterion also [29]. In other words, the test gives a lower probability of error of second kind than any other tests for the same or less probability of error of the first kind. If the distributions are Gaussian with variance-covariance matrix Σ_0 and mean vector $\underline{\mu}_0$ for class 0 and Σ_1 , $\underline{\mu}_1$ accordingly for class 1, then the likelihood ratio can be expressed in a more explicit form. Again n is the number of elements of vector \underline{x} and Σ_i^{-1} is the inverse matrix of Σ_i in the next equations.

As $f(\underline{x}/0) = (2\pi)^{-\frac{n}{2}} |\underline{\Sigma}_0|^{-1/2} \exp [-1/2(\underline{x}-\underline{\mu}_0)^T \underline{\Sigma}_0^{-1} (\underline{x}-\underline{\mu}_0)]$

and $f(\underline{x}/1) = (2\pi)^{-\frac{n}{2}} |\underline{\Sigma}_1|^{-1/2} \exp[-1/2(\underline{x}-\underline{\mu}_1)^T \underline{\Sigma}_1^{-1} (\underline{x}-\underline{\mu}_1)]$, then the likelihood ratio

$$L(\underline{x}) = (|\underline{\Sigma}_1|/|\underline{\Sigma}_0|)^{1/2} \exp(-1/2)[(\underline{x}-\underline{\mu}_0)^T \underline{\Sigma}_0^{-1} (\underline{x}-\underline{\mu}_0) - (\underline{x}-\underline{\mu}_1)^T \underline{\Sigma}_1^{-1} (\underline{x}-\underline{\mu}_1)]$$

The equation becomes more compact in form if we make $\underline{\Sigma}_0 = \underline{\Sigma}_1 = \underline{\Sigma}$ and by taking logarithms of both sides as

$$\ln L(\underline{x}) = -1/2[(\underline{x}-\underline{\mu}_0)^T \underline{\Sigma}^{-1} (\underline{x}-\underline{\mu}_0) - (\underline{x}-\underline{\mu}_1)^T \underline{\Sigma}^{-1} (\underline{x}-\underline{\mu}_1)]$$

Without utilizing the knowledge of quadratic form, the above expression can be simplified to a linear form as shown in Appendix A, to,

$$\ln L(\underline{x}) = \underline{x}^T \underline{\Sigma}^{-1} (\underline{\mu}_0 - \underline{\mu}_1) + \text{const.}$$

This is essentially a linear polynomial equation and of course easy to work with.

These quadratic forms represented as $Q(\underline{x}) = (\underline{x}-\underline{\mu})^T \underline{\Sigma}^{-1} (\underline{x}-\underline{\mu})$ imply the square of distance between \underline{x} and $\underline{\mu}$, and are optimal for the Laplace and rectangular distributions [26], as well as the aforementioned Gaussian distribution.

Going back to Bayes' decision, which requires minimum probability of error, it is understandable that a decision should be made to assign an unknown \underline{x} to the one of k classes for which $f(\underline{x}/k)$ is greater than any other classes. For those distributions stated above, decisions can be made by only comparing the quadratic form itself if there are some reasonable assumptions.

2. Learning with a teacher

When the functional form of conditional distribution with unknown parameters is given together with a set of samples from known classes, the given samples would be used as a training set to estimate the unknown parameters. Writing the sets of samples as $X^i(n) = \{\underline{x}^i(1), \dots, \underline{x}^i(n)\}$, $i = 0, 1$ and the conditional probability distribution functions of each class as $f(\underline{x}/X^i(n), i)$ instead of the form $f(\underline{x}/\theta, i)$ for known parameters θ , the principal quantity of likelihood ratio can be represented with the same format as before.

$$L(\underline{x}) = f(\underline{x}/X^0(n), 0) / f(\underline{x}/X^1(n), 1)$$

The basic operation is to calculate $f(\underline{x}/X^i(n), i)$ for each i and it is done by a recursive procedure in Appendix B, as

$$f(\underline{x}/X^i(n), i) = \int f(\underline{x}/\theta, i) f(\theta/X^i(n), i) d\theta$$

$$\text{and } f\{\theta/X^i(n)\} = [f\{\underline{x}^i(n)/\theta\} f\{\theta/X^{i-1}(n)\}] / [\int_{-\infty}^{\infty} f(\underline{x}^i(n)/\theta) f(\theta/X^{i-1}(n)) d\theta]$$

Here, the expression of $f(\theta/X^i(n))$ is used for simplicity instead of $f(\theta/X^i(n), i)$. From this recursive way, $f(\theta/X^i(n))$ can be calculated and used for the likelihood ratio test even though it in fact may be difficult to execute. If the distribution is assumed Gaussian, then there is a direct way of calculating the parameters [28].

3. Learning without a teacher

If the given set of samples are not predefined or classified, then the method discussed in the "learning with a teacher" scheme cannot be used without modification. This so called, "learning without a teacher" case is quite realistic, but the difficulties of handling data are enormous and one usually resorts to suboptimum solution rather

than the direct application of procedures.

Learning with or without a teacher method is not easy. In each stage of calculation of conditional probability density functions, the system should be capable of adapting itself for various operations, both linear and non-linear, and the possibility of this happening makes the predictions on system behavior very difficult. The realization of the system is complex, also.

4. When the functional forms are not known

This is the most general and difficult of the four cases. The data sets are given without any prior knowledge on the functional distribution, and the classification of samples may or may not be known. There is no conclusive result on the case when the samples are not classified [28], [38].

If the samples are from known classes, two deterministic approaches exist. The first one is to find a linear decision function which is valid at least for the given samples of known classification [35]. The assumption is that a sufficient number of samples are available. A linear classifier thus assigns an unknown pattern \underline{x} to class 0 if $\underline{x} \cdot \underline{w} > C$ and to class 1 otherwise. The coefficients w_j of \underline{w} are proportional to the components of a vector onto which the patterns are projected. The simplest method of computing the parameters of a linear classifier is to let $\underline{w} = \underline{S}^0 - \underline{S}^1$ where \underline{S}^i 's are typical members of the two classes. Quite often these \underline{S}^i 's are set equal to $\underline{\mu}_0, \underline{\mu}_1$, the mean vectors of the samples.

As the functional forms of the distributions are not given, then \underline{w} , which minimizes the error probability, cannot be solved analytically.

This deterministic method requires an optimization procedure to calculate the coefficients. Since this deterministic method does not make use of any a priori probability, it lacks the property of quantitative evaluation of the performance.

The second method achieves pattern recognition using a conditional probability density function $f(i/\underline{x})$ [30]. If the probability density function can be expanded into a series, then the decision function $g(\underline{x})$, which classifies a given set of data to class 0 if it is positive and class 1 if it is not, can be expressed as

$$\begin{aligned} g(\underline{x}) &= f(1/\underline{x}) - f(0/\underline{x}) \\ &= 2 f(1/\underline{x}) - 1 = \sum_{j=1}^{\infty} w_j g_j(\underline{x}). \end{aligned}$$

To determine g_j is another difficult problem and usually orthonormal functions are used. Suppose that g_j 's are defined, then the problem which remains is only to calculate w_j 's for values of the functions measured at random points.

5. Sequential decision methods

In the previous sections, certain satisfactory numbers of features or measurements were assumed to be fixed and every method was mentioned without asking the question, "How many measurements should one take from a class?" There should be at least enough features or measurements, but the number cannot be increased indefinitely because of the cost of taking measurements or the limitation in time [33], [34]. If the cost of taking measurements is significant or the features themselves are sequential in nature, then sequential methods should be used [8].

It is specifically important to have the data in such an order that the decision should be terminated at the earliest stage possible. After the n -th feature measurement is taken, the likelihood ratio

$$L(\underline{x})_n = \prod_{i=1}^n f(\underline{x}/H_0)/f(\underline{x}/H_1)$$

is calculated and compared with two stopping boundaries A and B .

If $L_n \geq A$, then \underline{x} is classified into class 0 and if $L_n \leq B$, then \underline{x} is classified into class 1, otherwise the same process is repeated for the $(n+1)$ th measurement.

The stopping boundaries A and B are set in much the same way as the threshold is determined in Neymann-Pearson criterion for fixed number of measurements, or

$$A = (1-\gamma)/\delta, \text{ and } B = \gamma/(1-\delta)$$

where γ and δ are set by the user.

It is shown, for a two-class decision problem, that a sequential decision method has an optimal property in the sense that it consumes the least number of features to make the same or lower probability of error compared to any other classification algorithms [9]. If the functional forms of distributions are not given, learning schemes should also be adopted in addition to the use of the sequential method.

The four cases cited before in connection with the available information about the distributions, and the possible algorithms that could be adopted for classifications are tabularized for simple display in Table II-1.

6. Summary of the parametric methods

The basic properties of Bayes' optimal decision rule are

discussed along with Neymann-Pearson criterion for the case when complete a priori information is known for the conditional distributions. If the parameters of the distributions are not given, the samples from known classifications can be used to estimate the parameters. If the sample classes are not given, a nonsupervised learning method is necessary. Every method mentioned can be substituted by sequential decision procedures which guarantee the optimal solution. When no functional form is supplied with samples, and this is the most probable case of all, deterministic ways of using discriminant functions or stochastic methods are available for substitution, but no absolutely general method is in existence.

While these parametric methods seem straightforward and mathematically eligible for further development of algorithms, it should be also noted that the assumptions set for the parametric methods do not always conform to practical situations. In fact, the functional form of distributions are not known and their forms are rarely Gaussian [1], or after non-linear transformations which are commonly used, the data certainly will not remain Gaussian if the original data are Gaussian [31]. The learning with or without a teacher is in most cases too involved and not easy to implement. The motivation to investigate nonparametric methods is thus aroused.

Table II-1. Available information and possible algorithms of the four cases

		Case 1		Case 2	Case 3	Case 4
Available information	Shape of distribution	known		known	unknown	unknown
	Parameters of the pdf	unknown		known	unknown	unknown
	Set of samples	given		not necessary	given	not given
	Sample classification	given	not given	not necessary	given	not given
Possible algorithms	Direct use of Bayes' rule	no	no	yes	no	no
	Learning with a teacher	yes	no	no	no	no
	Learning without a teacher	no	yes	no	no	no
	Deterministic method	possible	possible	possible	yes	yes
	Sequential decision	yes	yes	yes		

B. Nonparametric Methods

For most general situations in which little is known about the distribution of random variables, it is necessary to develop methods that do not depend on any particular form of a probability density function, or on less restrictions on the form of distributions. A nonparametric method can be used when less than a complete knowledge of the pdf is provided and the estimation of the distribution is impossible with a finite number of parameters.

The term "nonparametric" comes from the fact that these tests do not test or estimate the parameters of distributions as is done for parametric methods. Since this category of statistical methods requires very little knowledge of the distribution of the variables, the name "distribution-free method" is also often used.

Karl Pearson's chi-square test of fit [14] proposed in 1900 is one of the earliest nonparametric methods but relatively little concern was directed to this somewhat unfamiliar field of statistics until Wilcoxon's rank method was introduced in 1945. This test showed remarkable performance in its simplicity and relative error probability, even when the distributions are Gaussian. These nonparametric methods thus have advantages which are: (1) insensitivity to the input variables statistics while a fixed maximum error probability in one class is maintained, (2) relatively easy implementation of the system and software resulting in reduced time for calculation.

While the lack of statistical utilization of information about the input variables keeps one from designing an absolutely optimal system,

it should also be remembered that an optimum system is not always feasible in practice. Nonparametric methods are worth consideration. The performance figure of nonparametric methods is considered next.

1. Asymptotic relative efficiency

Asymptotic relative efficiency (ARE) is used as a figure of merit of one pattern recognition algorithm with respect to another method for the same hypothesis test.

Let N_1 and N_2 be the smallest number of observations needed for each of the two algorithms to be compared to reduce the β error at most below a certain value while maintaining the same fixed α error. Pitman's relative efficiency is defined as

$$e_{1,2} = N_1/N_2$$

This ratio should be a function of α , β and the probability density function of each class [2], [7], or

$$e_{1,2} = n_1(\alpha, \beta, f(\underline{x}/H_0), f(\underline{x}/H_1)) / n_2(\alpha, \beta, f(\underline{x}/H_0), f(\underline{x}/H_1))$$

As the relative efficiency defined above is difficult to evaluate for any arbitrary α , β , and pdf's, the asymptotic relative efficiency is derived for simplified comparison by letting N_1 and N_2 approach infinity. However, it is necessary to reduce the signal level to zero in order not to have β become zero with infinite number of samples as it would be for consistent statistics.

Then the ARE is,

$$ARE_{1,2} = \lim_{\substack{N_1, N_2 \rightarrow \infty \\ H_1 \rightarrow H_2}} e_{1,2}(\alpha, f(\underline{x}/H_0), f(\underline{x}/H_1), N_1, N_2) \Big|_{\beta=\beta_1}$$

The subscripts specify the ARE of method 2 compared to method 1. Allowing H_1 to approach H_0 is in analogy to taking a relative efficiency of two system performance in weak-signal condition, hence an ARE less than unity means that algorithm 2 is less efficient than algorithm 1. An ARE more than unity means that algorithm 2 is better than the other. It is true that ARE gives a measure of comparing two methods in performance, but its engineering value has not yet been completely proven.

2. One-input tests (With Reference Noise)

Suppose there is only one input channel and each measurement vector obtained from either of the two classes has data length of n . Several methods are available to process the data.

a. Sign test

This test is sensitive to the difference of the medians of the two classes provided one of the medians is at the origin.

Let H_0 , H_1 be the null and alternative hypothesis, respectively, and \underline{x} an input vector as before. If the x_i 's, the elements of \underline{x} , are all independent and identically distributed with the same cumulative distribution function $F(x_i)$, then the null hypothesis H_0 is that $F(0)$ equals one half and the alternative H_1 is that $F(0)$ is not equal to one half.

For class 0 the probability of positive observations occurring is the same as that of the negative observation occurrence. For class 1 with median values not equal to zero, the probability of observing positive or negative values is greater than that of the opposite sign observations. This test calculates the number of positive or negative observations, whichever is smaller, and compares it to a certain threshold.

If the observation number exceeds the threshold, H_0 is accepted, otherwise H_1 is accepted. The threshold is determined in the following way:

For the case when only class 0 is present, the probability of observing positive signs is the same as that of observing negative signs.

$$\text{Let } U_i \begin{cases} 1, & \text{if } x_i \geq 0 \\ 0, & \text{if } x_i < 0 \end{cases}$$

then U_i corresponds to a single independent variable with equal probability of occurrence of a 1 or 0. The sum, $m = \sum_{i=1}^n U_i$, will be binomially distributed corresponding to n independent trials of an experiment with equal a priori probabilities of negative and positive sign observations. Naturally the number of positive or negative signs will be changed appreciably from the mean value of class 0, or $n/2$, whenever a set of data from class 1 is processed. If the data in class 1 have more positive median than class 0 data, the number of positive observations will be greater than the number of negative signs. For class 1 with a more negative median, the reverse will be true. To a certain predetermined significance level, the number of positive or negative observations is compared and determined whether class 0 is present or not. For example, suppose a significance level of 10 per cent is selected. With class 1 which has the more positive median and for sample size $n = 12$, m should be less than 4 because

$$(1/2)^n \sum_{m=0}^3 \binom{12}{m} (1/2)^n < 0.1 < (1/2)^n \sum_{m=0}^4 \binom{12}{m} (1/2)^n$$

For a sufficiently large number of samples, the binomial distribution can be approximated by a normal distribution with mean $\mu_1=np=n/2$ and variance $\sigma_1^2=npq=n/4$ [22]. The threshold C is determined by

$F\left(\frac{C-n/2}{\sqrt{n/2}}\right) = 1 - \alpha$ where $F(x)$ is cumulative normal distribution function with $\mu=0$ and $\sigma^2=1$.

It has been proven that the sign detector has an ARE of about 64% compared to a linear optimal detector of dc signals in Gaussian noise. For a noise other than Gaussian, like the Laplace distributions, the efficiency becomes greater than unity [7].

b. Wilcoxon's signed-rank test

This test is also sensitive to the difference in the median between the classes and requires the class 0 distribution be symmetric about the origin. It is said that the nonsymmetry of class 1 can be detected through this test [1].

On the contrary to the sign test which does not use much information about the input data except for the signs, this test uses the information of ranks of each observation. This implies that the signs are weighted according to the distance from the origin. For this test, a set of data is ordered and ranked according to their absolute values in increasing order and one takes the rank sum of positive data. From H_0 , for which $F(0) = 1/2$ and with identical distribution for each observed variable, it is clear that the sum of ranks of positive or negative observations should be a random variable. Each of the 2^n sets of possible sums has the same probability of occurrence, so that the distribution of sums is predetermined. The distribution will range

from 0 to $n(n+1)/2$ with mean at $n(n+1)/4$. For the alternative H_1 , for which $F(0) < 1/2$ because of the positive signal, the number of positive observation ranks will be more than that of negative observation ranks. If a particular sum of ranks of an observation falls into a region within a certain threshold, then the hypothesis H_0 is accepted, otherwise, H_1 is accepted. Mathematically,

$$\text{if } \sum_{i=1}^n d_i \geq C, \text{ accept } H_0 \text{ or}$$

$$\text{if } \sum_{i=1}^n d_i < C, \text{ accept } H_1$$

$$\text{where } d_i = \begin{cases} 0, & \text{if } x_i < 0 \\ i, & \text{if } x_i > 0 \end{cases}$$

The threshold is obtained by the direct use of α error, the error probability of type 1, since a fixed number m of 2^n combinations should be outside of the threshold, or $m/2^n \leq \alpha$.

For a sufficiently large number of observations, the distribution of rank sums can be approximated by normal distribution [16] with mean $\mu = n(n+1)/4$ and variance $\sigma^2 = n(n+1)(2n+1)/24$, hence threshold C is calculated from

$$F[(C - \mu)/\sigma] = 1 - \alpha$$

It is found that the Wilcoxon's signed-rank method has an ARE of about 95.5 per cent with respect to optimal linear detector if the distribution is Gaussian but it increases considerably to more than 100 per cent as the distribution is drifting away from Gaussian [7]. A more impressive result was reported by Hodges and Lehmann [15]. They showed that the ARE of the Wilcoxon's test relative to the t-test is

never less than 86.4 per cent for any kind of distributions $F(x)$, and is arbitrarily high without bound. So the linear test requires only 13.6 per cent less data than Wilcoxon's test at its best for the same performance, but it may require more samples in many cases.

c. Sequential nonparametric method

Sequential methods in parametric cases are mentioned in part A of this chapter. According to Fu and Chien [9], significant findings were made in recent years in calculations of sequential distributions and the practical use of it to nonparametric case. In applying ordinary sequential probability ratio test to its nonparametric cases, it is necessary to find out the probability distributions of the sequential rank vectors $\underline{r}(n) = (r_1, r_2, \dots, r_n)$ of original vector \underline{x} . The sequential rank is represented as r_n if x_n is the r_n -th smallest element in the sample vector \underline{x} .

Since there exists a one-to-one correspondence between the ordered observations and the sequential rank vector, the distribution of the sequential rank is completely determined by the ordered observation. If x_i 's are all independent, then,

$$\begin{aligned} F\{\underline{r}(n)\} &= F(x_1 \leq x_2 \leq \dots \leq x_n) \\ &= \int \dots \dots \dots \int \prod_{j=1}^n dF_j(x_j) \\ &\quad -\infty < x_1 \leq \dots \leq x_n < \infty \end{aligned}$$

where $F_j(x_j)$ indicates the distribution functions of x_j .

If Lehmann's alternatives are adopted for the distribution functions $F_j(x_j)$, then

$$F_j(x_j) = F^{r_j}_j(x_j) = \{F(x_j)\}_j^{r_j}, \quad r_j > 0$$

where r_j is the observed sequential rank,

$$\begin{aligned} \text{or } dF_j(x_j) &= dF^{r_j}_j(x_j) \\ &= r_j F^{r_j-1}_j(x_j) \end{aligned}$$

From this,

$$\begin{aligned} F(x_1 \leq x_2 \leq \dots \leq x_n) &= \int_{-\infty < x_1 < \dots < x_n < \infty} \dots \int_{j=1}^n dF^{r_j}_j(x_j) \\ &= \prod_{j=1}^n r_j / \prod_{j=1}^n \left(\sum_{k=1}^j r_k \right) \end{aligned}$$

which is found by some simple manipulation.

Relabeling the x_j 's, the probability of any order of the x_j 's can be determined.

3. Two-input tests

Suppose that, in addition to the channel of the one input case, there is another statistically independent noise channel which is not perturbed by the presence of signals. Let this additional set of reference noise input data by $\underline{y} = (y_1, y_2, \dots, y_n)$ which is independent of \underline{x} . This situation should not be confused with the case where the presence of signal changes statistics in both channels. The latter case is mentioned in later section.

With the same assumptions made in one-input test, the null hypothesis is that the median of one class is the same as that of the other against the alternative that the medians are different, or

$$H_0 : F(z_i = 0) = 1/2$$

$$H_1 : F(z_i = 0) \neq 1/2$$

where $z_i = x_i - y_i$.

The same procedures discussed in previous section of one-input case can be employed by treating z_i as the variable x_i of one-input case.

a. Sign test

This test calculates the number of positive or negative signs of $z = x - y$ and compares the number to a certain threshold. As in the one-input sign test, the threshold is found from the fact that signs of observations are elements of a random vector which has equal probability of occurrence of either positive or negative signs if H_0 is true. For a sequence of random signs, the distributions should be binomial. If a distribution falls beyond a predetermined threshold of the binomial distribution, H_0 is rejected, otherwise H_1 is rejected.

b. Wilcoxon's signed-rank test

Like the sign test for the two-input case, this test also makes use of the same concept as for the one input case. First, determine the signed differences of the two sets of observations, x and y . Let z be the signed differences in vector form. Then determine the ranks of elements according to their absolute values. These ranks are then attached with positive or negative signs, which ever are original. For an alternative hypothesis which has more positive median than the null hypothesis, there will be more positive elements than negative ones, hence more positive ranks than negative. The next step is to

find the sum of ranks of positive signs if the alternative has more positive median, negative signs if the alternative has more negative median.

Since the sum of ranks is a random variable which is approximately normally distributed with mean $\mu = n(n+1)/4$ and standard deviation $\sigma = \sqrt{n(n+1)(2n+1)/24}$ [16], the probability of a value of sum as extreme as it can be computed. If it falls beyond a threshold, that the distribution differs distinctively, the alternative H_1 is accepted. Otherwise H_0 is accepted. The above can be expressed simply as,

$$\sum_{j=1}^n \sum_{i=n}^n U(x_i - y_i) \begin{cases} \geq C \rightarrow H_0 \text{ is accepted} \\ < C \rightarrow H_1 \text{ is accepted,} \end{cases}$$

for positive alternative. Threshold C is determined in the same way as for the one input case.

c. Rank-sum test

It has been shown that the previously mentioned sign or signed-rank tests require some restrictions on the distribution shapes. For the test of the null hypothesis H_0 , that the two \underline{x} and \underline{y} are from identical distributions, against the alternative H_1 , that the two are from different distributions, the Wilcoxon, Mann-Whitney's rank-sum method can be used. It is assumed that the x_i 's are independent and identically distributed as all y_i 's are, also.

Different sample sizes may well be used, so let m and n be the number of measurements in each of \underline{x} and \underline{y} , respectively. From the assumption of independent random measurements of x_i 's and y_i 's, each of the $(n+m)!$ possible permutations of measurements of the original sets must have the same a priori probability if the null hypothesis of

identical distribution is true. In other words, any of the $\binom{m+n}{m}$ possible combinations of \underline{x} and \underline{y} data sets from $(m+n)$ measurements were equally probable to have become the actual observation set.

For each of these $\binom{m+n}{m}$ possible data sets, there exists a value $\sum_{i=1}^m R_i$, where R_i is the rank of x_i in size among the $(m+n)$ observations. Then the distribution of $\sum_{i=1}^m R_i$ values of all the $\binom{m+n}{m}$ possible data sets must conform to a predefined distribution to satisfy H_0 . The null hypothesis of identical distribution is rejected if the actual values of $\sum_{i=1}^m R_i$ falls outside of preselected significance level. This test is very sensitive to the difference in the level (mean) separation but is also somewhat sensitive to the difference in shape and variance [1].

The ARE of the rank-sum test with respect to the optimal linear detector for Gaussian distribution case is 0.955. This is the same as that of signed-rank test since both methods are the same for symmetrical distributions. The ARE of this test never falls below 0.864 with respect to the optimal linear detector and can be arbitrary high for many distributions [15].

4. Two-input tests (correlation method)

Assume a system with two input channels which have statistically independent noises but the presence of a signal perturbs both channels simultaneously. The appropriate test decides on the hypothesis that the two channels are independent versus the alternative of dependence. This kind of situation occurs in the practical case of the scattered or fading radio communication channel. The two most widely used non-parametric methods for testing if correlation exists

are the rank correlation and polarity coincidence correlation methods [6], [22].

a. Rank correlation method

Let each of the two channels be represented by \underline{x} and \underline{y} , respectively. A pair (x_i, y_i) is a sample which is obtained at the same instance of observation, or matched observations. The ordinary linear sample correlation coefficient is defined as

$$r = \frac{\sum(x_i y_i) - (1/n) \sum x_i \sum y_i}{(\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2)^{1/2}}$$

Nonparametric rank correlation coefficient is found by the same procedures except that the actual values of x_i and y_i are replaced by their respective ranks among each \underline{x} and \underline{y} .

This method is also called Spearman's rank correlation test. If the coefficient is less than a predetermined threshold C , the hypothesis H_0 that the two channels are not correlated is accepted. If r exceeds C , then the alternative H_1 of dependence is accepted. The ARE of this rank correlation technique is known to be 0.91 with respect to ordinary linear sample correlation methods if the sample distributions are Gaussian. The ARE can be greater than unity if the distributions are not Gaussian [1].

b. Polarity coincidence correlation

If only the polarities of each sample-pair are examined for a test, the least complicated method is available. The total number of points (x_i, y_i) which fall in the first and third quadrants of the x - y plane can be written in the form $\sum_{i=1}^n U(x_i, y_i)$.

This polarity coincidence correlator decides that a signal is present when the above value exceeds a threshold. The ARE of this test is shown to be 0.202 for Gaussian distribution with respect to optimal detector. As usual, the threshold should be set at an appropriate level which conforms to significant correlation between the two channels.

5. Briefs on nonparametric methods

Nonparametric methods for the one and two input cases were reviewed. Even though the practical usefulness of ARE is not yet thoroughly investigated, the nonparametric methods have very good relative efficiencies for distributions other than Gaussian. For nonparametric methods, the probability of making an error of one kind can be preset to a value no matter what distribution forms the random variable has, and just a few general assumptions are necessary to proceed. The assumptions are: (1) continuous distribution over a range of the variable, (2) different median of each class, and (3) symmetrical distribution for signed-rank test.

The sequential method has also been considered. The sequential distribution of ranks has a one-to-one correspondence with the ordered measurements. Hence, assuming the Lehmann's alternative, the probability of any order of sequential rank vector can be calculated.

CHAPTER III

BASIC PERFORMANCE COMPARISONS

The comparisons of the performance of parametric and nonparametric methods described earlier are made in this chapter. For ease in analysis of these methods the performance of the Bayes' classifier is used as a reference. Random variables of known probability density functions are used as corrupting noise. The pdf's used are the Gaussian, two-sided exponential or Laplacian and the Rayleigh distributions.

Only the two-class problem is investigated using the algorithms discussed in Chapter II. The generalization of the two-class problem into a multi-class one is done in Chapter IV. The signed-rank and sign tests are employed extensively. K-class algorithm [39], which is one of the nonparametric methods developed recently, is also used. The use of the K-class algorithm was made possible by a subroutine supplied by G. Nelson of the Electrical Engineering Department and the Remote Sensing Institute of South Dakota State University.

To perform computer simulations of the different methods, a random sequence of signals and random noises of known distributions were generated according to the procedures discussed next.

A. Generations of Random Signals and Noises

1. Random signals

The computer subroutine RANDU is used to generate 512 uniformly distributed random variables from 0 to 1. The reason for choosing this 512 is that it is large enough to give consistent error probabilities for each algorithm and is not too large to process by computer.

Since this is only a two-class problem, signal zero is assigned if the uniform random variable has a value lower than one-half and signal one is assigned if the variable is greater than one-half. Because the distributions are uniform, the a priori probabilities of signal zero and one occurring are equal to one-half. In analogy to the communication's problem, zero may represent that there is not a signal present, while a one indicates the presence of a signal with unit amplitude. For each signal of zero or one, sixteen samples are taken and corrupted by independent noises. The problem is to determine whether the signal was originally zero or one, using different algorithms. When Laplacian and Rayleigh distribution noises are used, sample sizes of four and eight are used additionally to investigate the effects of the sample sizes on the probability of error.

2. Random noises

Three general approaches to numerical generation of random variables with a given distribution are available. The so-called inverse transform technique is the easiest one to work with if the cumulative distribution function $F(x)$ of the random variable is known. Since any cumulative distribution function is defined over the range of zero to one and a uniformly distributed random variable r can be generated over the same range by using the subroutine RANDU, r may be set to equal to $F(x)$. For every r there is a unique x which is calculated by taking the inverse transform of the cumulative distribution function or $x = F^{-1}(r)$. As r is a uniform random variable and $f(x)$ is the derivative of $F(x)$, x is the desired value of the random

variable with the specified pdf $f(x)$.

Mathematically,

$$r = F(x) = \int_{-\infty}^x f(x) dx$$

$$\text{and } F(x) = p(x' \leq x) = p[r \leq F(x)] = p[F^{-1}(r) \leq x]$$

hence, $x = F^{-1}(r)$ is the random variable with density function of $f(x)$.

The above procedure is applied very easily to a two-sided exponential distribution $f(x) = (c/2) \exp(-c |x|)$ whose cumulative distribution function is

$$F(x) = \begin{cases} 1 - 1/2 \exp(-cx) & \text{if } x \geq 0 \\ 1/2 \exp(cx) & \text{if } x < 0 \end{cases}$$

Since positive values of x correspond to $0.5 \leq r \leq 1.0$, and negative x to $0 \leq r < 0.5$, x is determined from each r as

$$r = 1 - 1/2 \exp(-cx) \rightarrow x = (-1/c) \ln(2 - 2r) \quad \text{for } 0.5 \leq r \leq 1.0$$

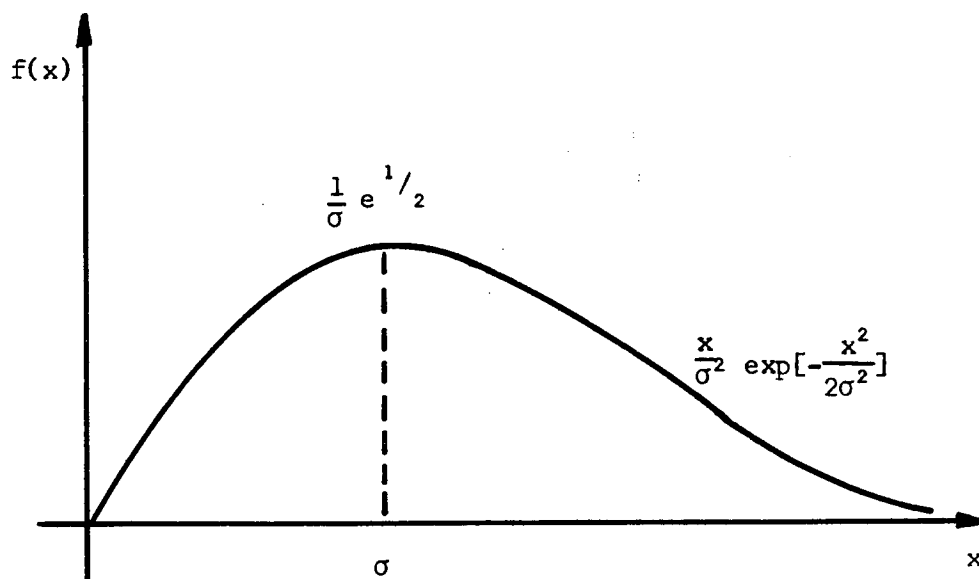
$$\text{and } r = 1/2 \exp(cx) \rightarrow x = (1/c) \ln(2r) \quad \text{for } 0 \leq r < 0.5$$

Random variable with Rayleigh distribution can also be found through the same procedures. The density function has the form

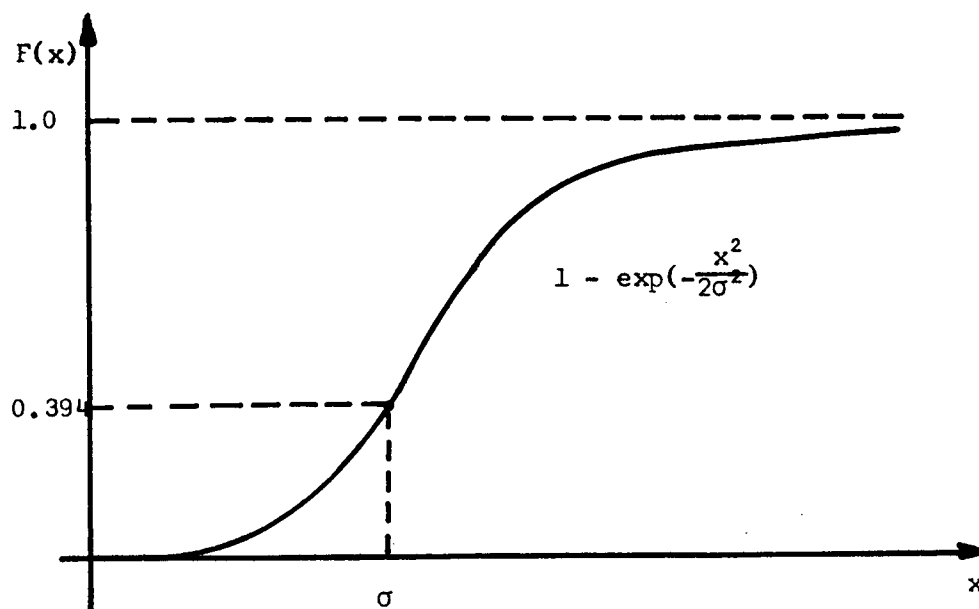
$f(x) = (x/\sigma^2) \exp(-x^2/2\sigma^2)$, $x \geq 0$ and the cumulative distribution function is $F(x) = 1 - \exp(-x^2/2\sigma^2)$ as seen in Figure 3-1. Since there is a unique x for every random variable r with uniform distribution over the range zero to one such that $r = 1 - \exp(-x^2/2\sigma^2)$, then

$$x = \{2\sigma^2 \ln(1-r)\}^{1/2}$$

Gaussian random variables are generated by use of the subprogram GAUSS which utilizes the central limit theorem with twelve variables which are independent and identically distributed. In this subroutine subprogram, random variables generated by RANDU with uniform



a. Density Function



b. Cumulative Distribution Function

Figure 3-7. Rayleigh distribution

distribution from zero to one are used.

B. Gaussian Distribution Case

The first simulation problem is executed with Gaussian noise case. To check the effect of the signal-to-noise ratio, five separate experiments with different mean values of signal one were used with a fixed value of variance. In other words, the distribution of signal zero has a mean value of zero and variance of one while signal one has mean values of 0.3, 0.5, 0.75, 1.0 and 1.25, respectively, for different experiments with the same variance. To each of 16 fixed mean values from whether a signal zero or one, statistically independent Gaussian noise with mean zero and variance one is added. Hence, for sample one the mean value increases to a value larger than zero. The case of correlated noise might have been studied here but it is avoided to concentrate only on the problem of comparing nonparametric methods to the parametric methods.

As the noise distributions are all independent and identical with each other, the optimal Bayes' decision is achieved by taking the sum of 16 observed values and comparing it to a threshold, which is determined by the following way. If the risk for making a decision in one class is the same as that of the other class, and the a priori probability is also the same in both classes, then the threshold is found from

$$\frac{f(\underline{x}/H_0)}{f(\underline{x}/H_1)} = 1, \text{ or } \prod_{i=1}^{16} f(x_i/H_0) = \prod_{i=1}^{16} f(x_i/H_1)$$

With independent and identical distribution for each variable, the

above can be reduced to linear form as in Chapter II, or

$x_1 + x_2 + \dots + x_{16} = 8.0$ if the signal one has mean value of one. Hence, for the sum which is less than eight, signal zero is assigned, otherwise signal one is assigned.

The nonparametric signed-rank test and sign test for the one-input case are also applied. As it was seen in Chapter II, the positive or negative rank sum for signed-rank test is a random variable whose range is from 0 to $136 = 16(16+1)/2$. This is readily understood because there are sixteen ranks from 1 to 16 according to the absolute values of observations and the signs attached to the ranks are the signs of original observations which are random in character. When signal zero is present, there will be almost equal probabilities of observing either negative or positive signs. But when signal one is present, the probability of observing positive signs will increase in accordance with the increase of mean value, making the sum of positive ranks more than that of the negative ranks.

For a large number of samples, say $n \geq 12$, distribution of signed-rank sums for signal zero can be approximated by Gaussian with mean $\mu = n(n+1)/4$, and variance $\sigma^2 = n(n+1)(2n+1)/24$. For the sample size 16 used in this experiment, $\mu = 68$ and $\sigma^2 = 374$. The threshold for this test is determined next. As an example, to make the α -error probability less than 5 percent, which is also the significance level of the hypothesis testing that a distribution is significantly different from the null distribution, the threshold C should be such that $F(C) = 0.950$. $F(x)$ is the cumulative distribution function of Gaussian case.

From the table of the cumulative distribution function of Gaussian pdf the following values were determined.

$$F(z) = 0.950 \rightarrow z = 1.64 \text{ where } \mu = 0 \text{ and } \sigma^2 = 1$$

To calculate the value of the threshold C , set $(C - \mu)/\sigma$ equal to 1.64. Then the value of C is 99.6.

Since the rank sum is an integer variable, the C should also take the form of integer. The nearest integer number to make the specified α -error probability is 100, which is the threshold value C . If the rank sum of positive signs is equal to or less than 100, signal zero is assumed to be present within 5 per cent of error probability. If the signed-rank sum is more than 100, signal one is present.

The sign test provides an easier arithmetic manipulation than the signed-rank test. Since the distribution of the number of positive signs or negative signs for signal zero is binomial, a threshold can be found from a binomial distribution table or by using Gaussian approximations for a large number of samples. For the approximation, the mean value is determined as $\mu = np$ and the variance as $\sigma^2 = npq$.

The number of either sign has only an integer value which ranges from 0 to n and the threshold is also discrete within this range. There are only $(n+1)$ possible threshold values. The threshold cannot be adjusted to a value which is a non-integer number to make significance level of the test arbitrary. The threshold for each signal level is determined according to the criteria discussed, and their values are given in Table III-1.

Table III-1 Thresholds for each test

Signal level Algorithm	0.3	0.5	0.75	1.0	1.25
Bayes' Optimal	2.4	4.0	6.0	8.0	10.0
Signed-rank	78.0	85.0	93.0	103.0	108.0
Sign test	9.0	9.0	10.0	10.0	10.0

The significance levels or α -error probabilities of the signed-rank test are set approximately at those of the calculated values of Bayes' classifier because the two classifiers are expected to perform equally well. This is expected since the ARE of the signed-rank test with respect to the Bayes is nearly one. For the sign test, the error probabilities are set at the nearest higher discrete value above the error probabilities of signed-rank test with the same conditions since the sign sum has only a discrete integer value from 0 to n .

The results of the computer simulation experiment are shown in Figure 3-2. As it was expected, the signed-rank test compares very well over the selected value of the mean difference between signal and noise. It works better than the optimum Bayes' classifier for the mean differences less than 0.75 and deteriorates a little beyond the mean differences of one. This degradation of performance may be from the fact that the α -error probabilities are predetermined and the error probabilities do not change no matter which distribution condition is used. The sign test has about five per cent more error probabilities than the Bayes' result, but it still performs well. The K-class

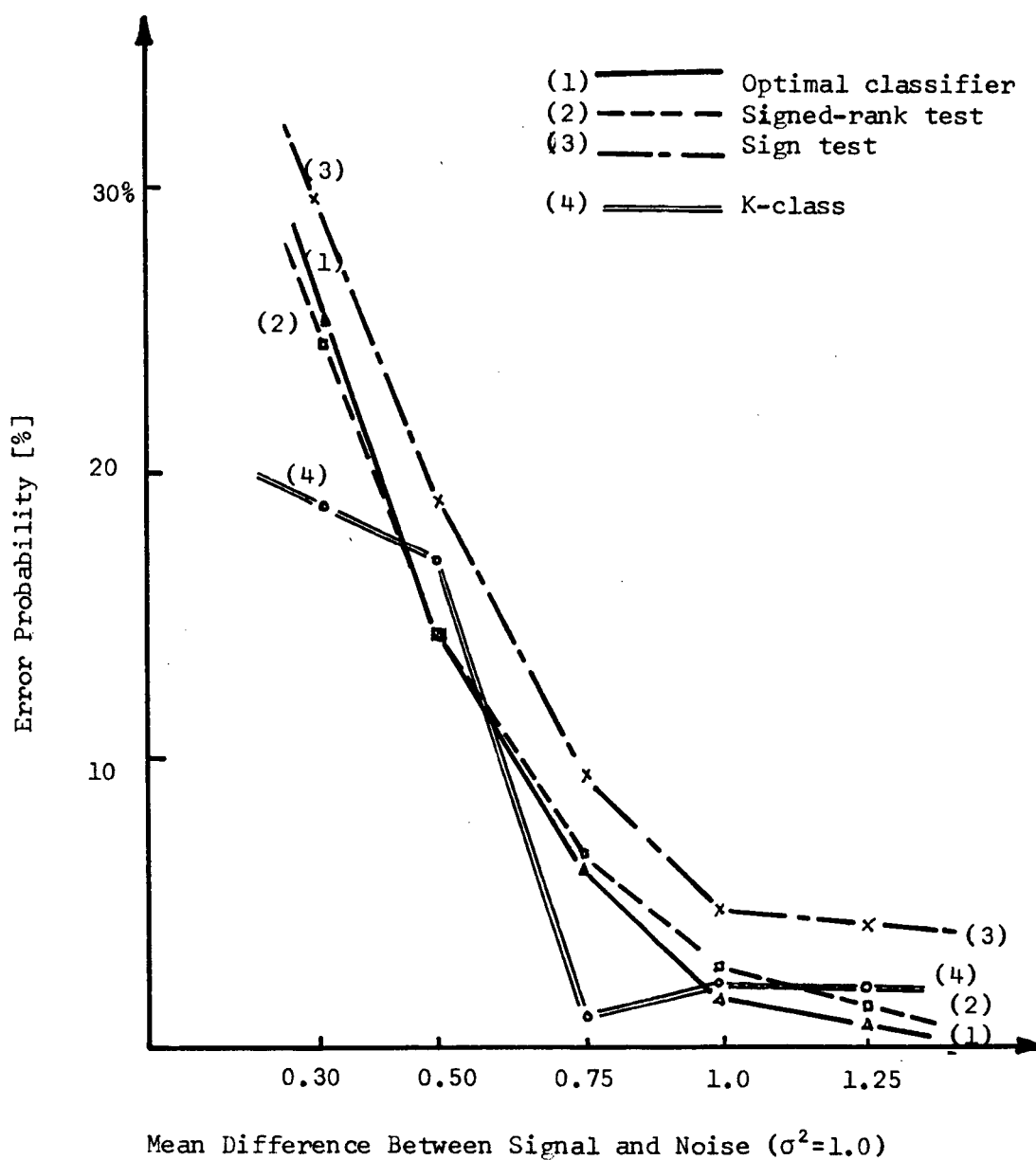


Figure 3-2. Error probabilities of different algorithms with Gaussian distribution (16 features).

algorithm performed exceedingly well. The reason seems to be that the test finds the optimized linear decision boundary without regarding any statistical distributions and that the complete data sets generated are used to train the algorithm. Additional computer simulation experiments on data not used to train the K-class classifier are necessary. It is important to note that the noise distributions simulated by computer are not pure Gaussian because only twelve uniformly distributed random variables are used to give a Gaussian variate.

The error probability ratios of Bayes' algorithm with respect to the nonparametric methods are given in Figure 3-3. The ratio is not the direct value of ARE but it gives the idea of how the nonparametric methods are working for different mean values. The figure shows that nonparametric method is more useful for small signal-to-noise ratio less than one. Since the absolute value of error probabilities for mean values greater than one is very small for either the signed-rank method or the sign test, the deteriorations of error probability curves do not necessarily mean that the nonparametric methods are impracticable.

To check the validity of these simulation experiments, theoretical error probabilities for Gaussian distribution case are calculated and compared to the values obtained from the experiment. When the a priori probability of each signal occurrence is equal to that of the other signal and each signal is uncorrelated, the average probability of error is, for Bayes' classifier,

$$P_e = 1/2[1 - \text{erf}(\mu/2\sigma)]$$

which can be readily calculated by use of a table or by computer program written to calculate the probability of error. These calculated

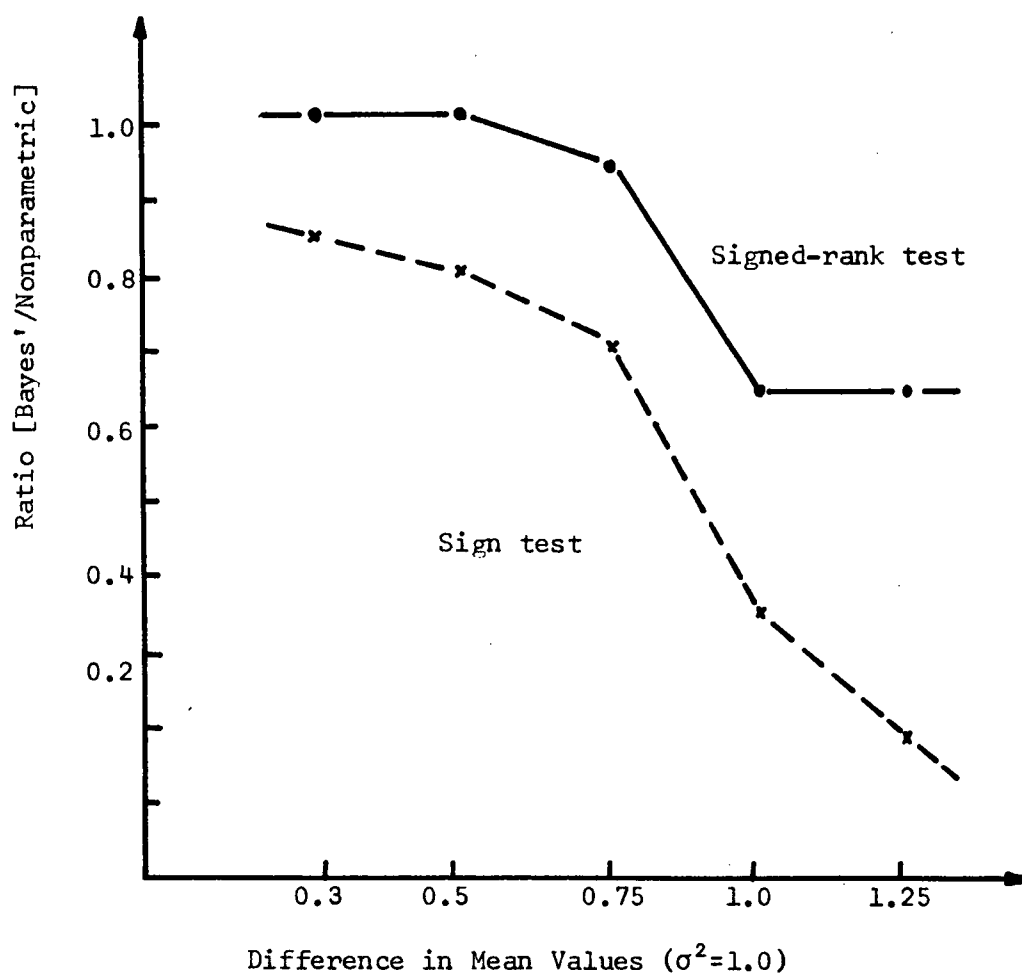


Figure 3-3. The error probability ratios between Bayes' and nonparametric methods [1.0 for Bayes']

values are compared to the experimental values in Figure 3-4. Since differences of these two sets of error probabilities are in the range of less than 3 per cent, it seems to be a reasonable conclusion that the simulation experiments are quite practical to evaluate the performance of these algorithms, so far as the Gaussian distribution is concerned.

Another point can also be mentioned. As it was noted before, a nonparametric method does predetermine the error probability of any one class. Table III-2 shows the predetermined and the resultant experimental values of the α -error probabilities in this simulation problem. It seems to be a general guide line to set the error probability of a class at about the same or a little higher value than that of Bayes' optimal classifier if it is known. This is because the performance of these methods is very close to each other. It is also found that the overall probability of error is very much affected by the value of the predetermined error probability of one class or the significance level. This is also observed in other distribution cases. Experiments on this phenomenon are performed with the Rayleigh distribution case.

Table III-2. Predetermined and experimental α -error

Mean value of signal		0.3	0.5	0.75	1.0	1.25
Algorithms		0.3	0.5	0.75	1.0	1.25
Signed-rank test	Predetermined	30	20	10	3.5	2.0
	Experimental Result	28.2	16.7	6.0	3.2	2.0
Sign test	Predetermined	about 20	about 20	about 10	about 10	about 10
	Experimental Result	19.1	19.1	8.7	8.7	8.7

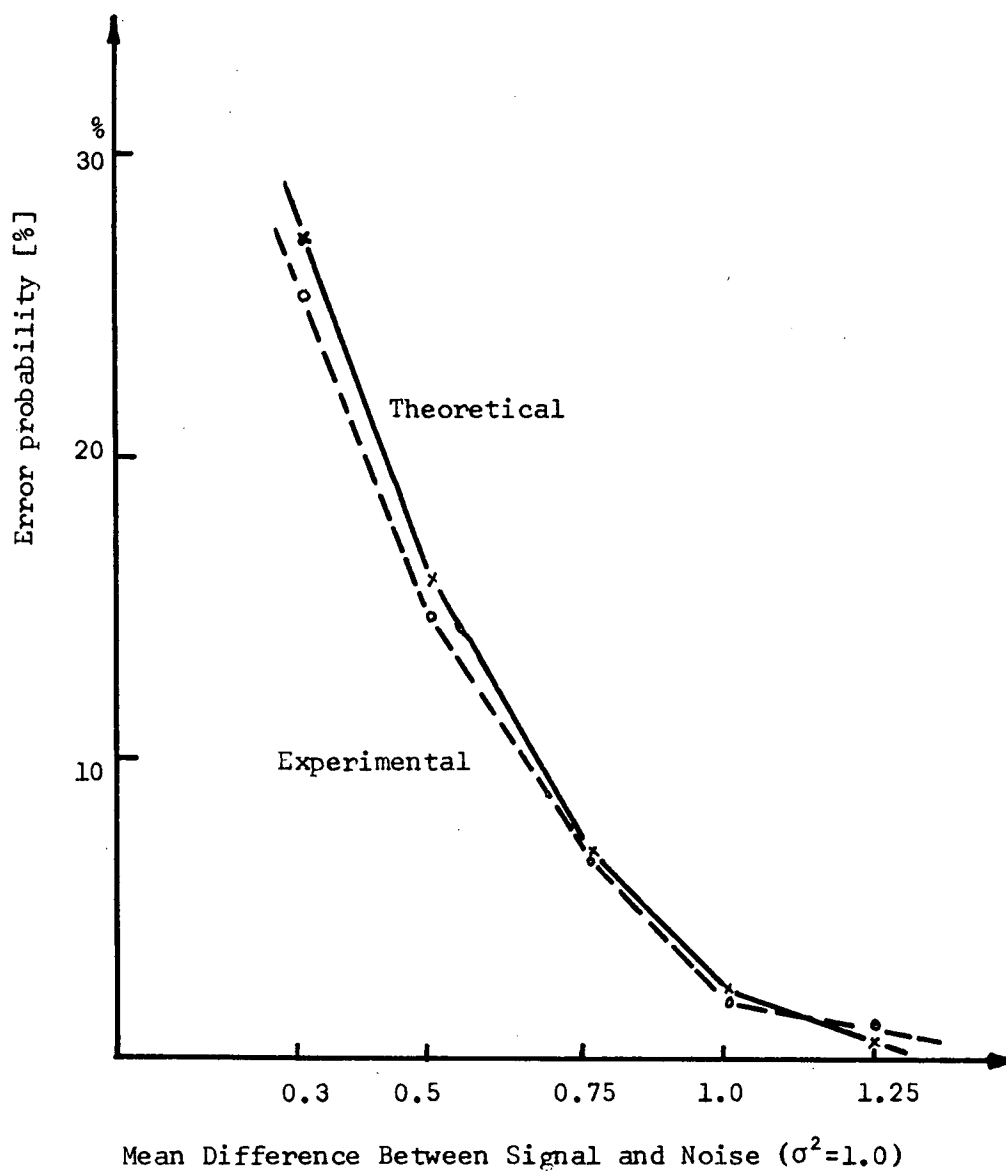


Figure 3-4. Theoretical and experimental error probabilities of Gaussian distribution (16 features)

C. Two-sided Exponential Distribution Case

Following the same procedures which are described in Chapter II, the two-sided exponentially distributed noises are generated and added to a random sequence of signal zero and one. Three different sample sizes of $n=4, 8$ and 16 are used for the five different mean values of signal one. The mean value of signal zero is always fixed to the value of zero. The five mean values of signal one are set equal to those of the Gaussian distribution case. The threshold for Bayes' optimal detector is determined by

$$\frac{f(\underline{x}/H_0)}{f(\underline{x}/H_1)} = \frac{\prod_{i=1}^n f(x_i/H_0)}{\prod_{i=1}^n f(x_i/H_1)} = 1$$

for independent and identical distribution of each random variable.

A priori probabilities and risks for making decisions are equal for both signals. Or

$$\exp[-k\{|x_1|+|x_2|+\dots+|x_n|-(|x_1-\mu_1|+|x_2-\mu_1|+\dots+|x_n-\mu_1|)\}] = 1$$

where k is a constant.

The above is reduced to

$$|x_1|+|x_2|+\dots+|x_n|-(|x_1-1|+|x_2-1|+\dots+|x_n-1|) = 0$$

where the mean value of signal one is one. For a given set of data \underline{x} , if the above calculation exceeds zero, signal one is decided, otherwise signal zero is decided.

Sign test and signed-rank test are applied as in the Gaussian noise case. Only the predetermined error probabilities of class one are set at a little higher value than that of Gaussian noise since greater

error probabilities are expected because of the distribution shape. In addition to the previous algorithms, the K-class algorithm and a classifier which operates with the assumption that the distributions are Gaussian are used.

The thresholds according to the number of samples and mean differences are calculated for different algorithms and are listed in Table III-3. Of course, random variables such as signed-rank sums or sign sums take on integer values only, but they are written in real type for use in computer programs.

It is interesting to note that there are only five possible thresholds for sample size of four, and nine possible thresholds to choose from for a sample size of eight, in the sign test. Only $(n+1)$ integer values are available for threshold values for sample size of n .

Table III-3. Threshold for each sample size and mean difference

Signal level		0.3	0.5	0.7	1.0	1.25
Algorithm	sample size					
Bayes' decision with Gaussian assumption	4	0.6	1.0	1.5	2.0	2.5
	8	1.2	2.0	3.0	4.0	5.0
	16	2.4	4.0	6.0	8.0	10.0
Signed-rank method	4	5.0	5.0	6.0	6.0	7.0
	8	19.7	21.5	23.7	25.7	26.7
	16	75.6	81.3	88.4	94.0	100.0
Sign test	4	3.0	3.0	3.0	3.0	4.0
	8	5.0	5.0	5.0	6.0	6.0
	16	9.0	9.0	10.0	10.0	10.0

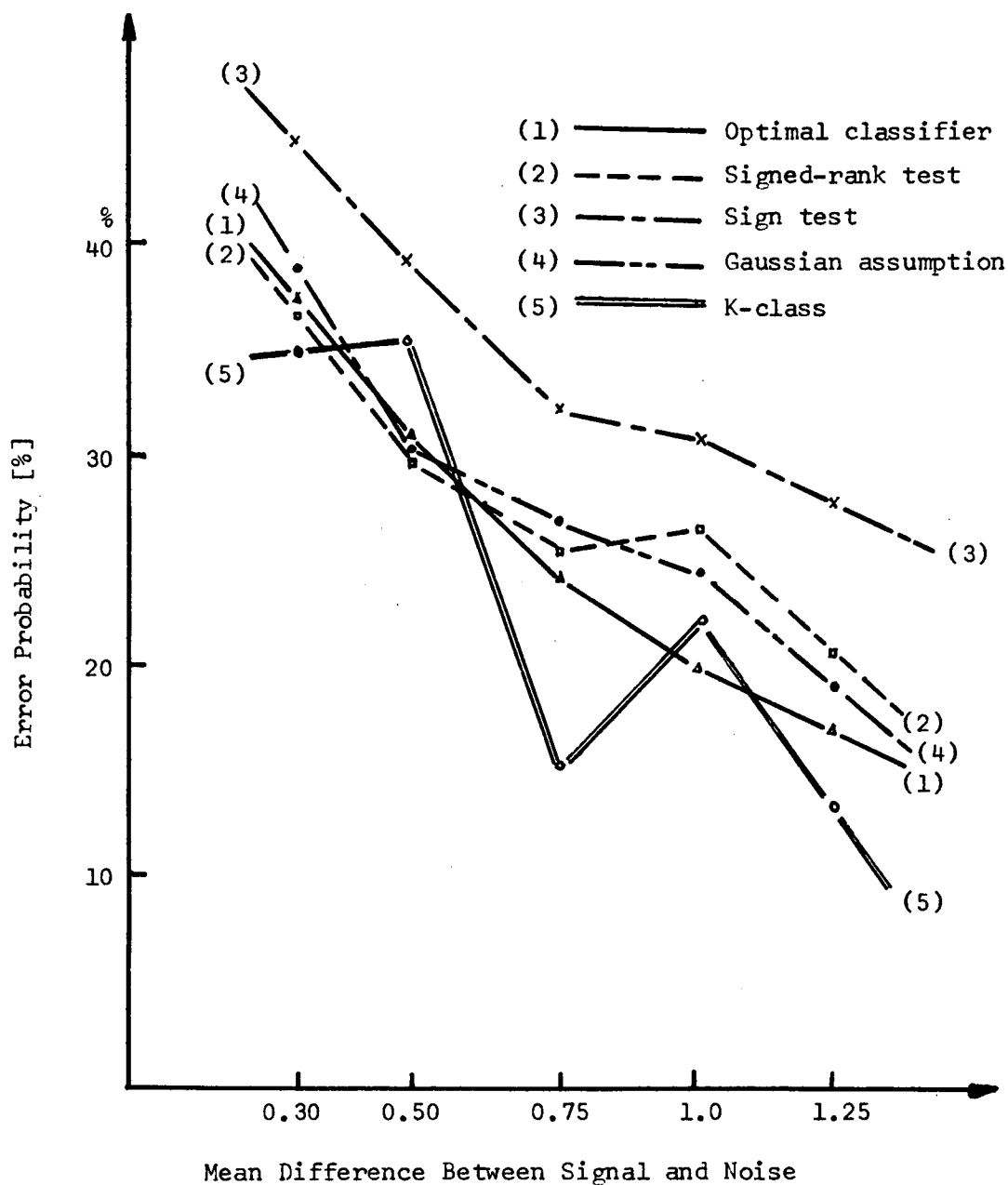


Figure 3.5.a. Error probabilities of different algorithms with Laplacian distribution (4 features)

Using the thresholds shown in Table III-3, performance tests are executed and the results are shown in Figures 3-5, a, b, c. In these experiments the Bayes' optimal classifier performed best as it should do. The other algorithms are close competitors. The signed-rank method proved to be better than any other algorithm except the Bayes'. The average error probabilities of each method for different sample sizes are given in Table III-4.

Table III-4. Average error probabilities for each algorithm with different sample sizes.

Algorithms Sample size	Gaussian Assumption	Signed-rank Method	Sign Test	Bayes'
4	0.2783	0.2773	0.3754	0.2579
8	0.2335	0.2265	0.2511	0.1939
16	0.1621	0.1466	0.1544	0.1201

Because the K-class algorithm performs with irregularity in error probability for different conditions of data, the average of the whole may not give much meaning, hence is omitted in the table. The expectation that the signed-rank test performs better than the algorithm using the Gaussian assumption is justified for sample sizes larger than four and mean difference less than one. It implies that the relative efficiency is more than unity for the nonparametric signed-rank test compared to the linear classifier; an agreement with the ARE value which is more than unity for the two-sided exponential distribution case.

The sign test seems to be too difficult for small sample sizes. However, for sample sizes of eight or more, it works almost as well

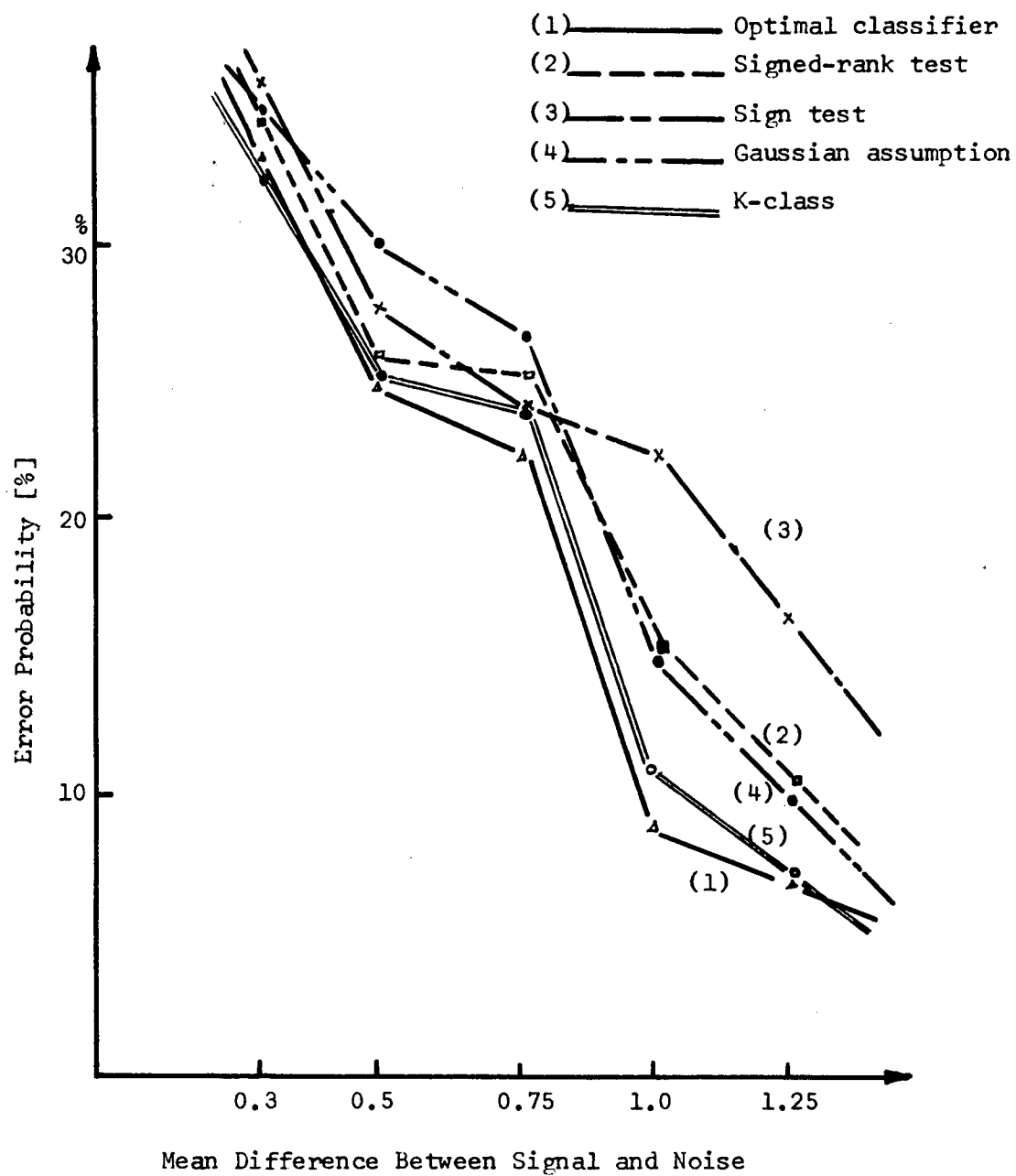


Figure 3-5.b. Error probabilities of different algorithms with Laplacian distribution (8 features)

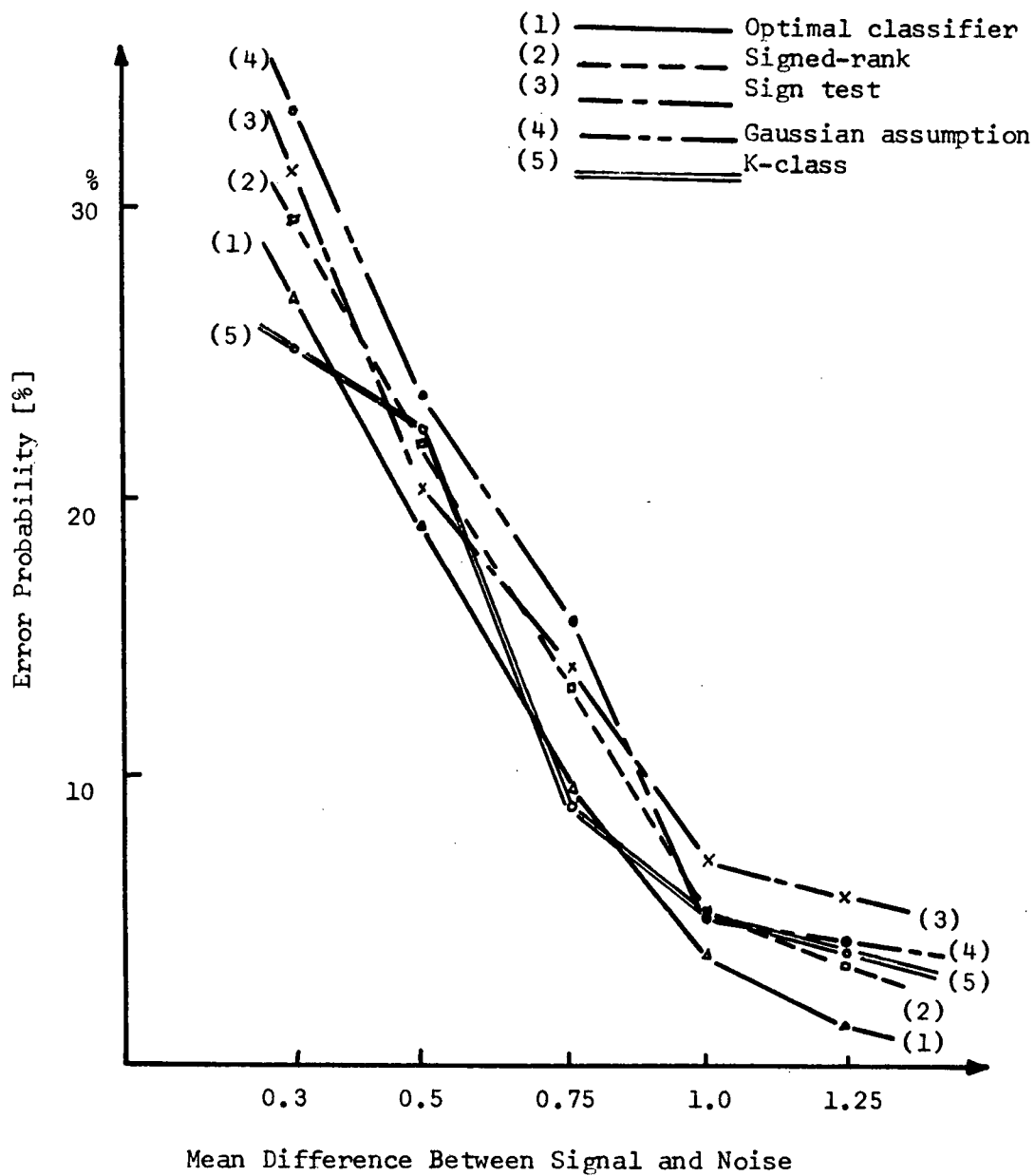


Figure 3-5.c. Error probabilities of different algorithms with Laplacian distribution (16 features)

as the others do. Some of the nonparametric thresholds could have been adjusted to more appropriate values which give less probabilities of error.

The K-class algorithm performed very well, again. It works better than optimal classifiers in some occasions but with much more fluctuation in error probabilities for different conditions. This irregularity in performance is excessive for small sample size. The same fact was seen in Gaussian and Rayleigh distribution cases. One of the reasons is that the relatively small number of signals are used. Instead of the 512 signals used for other algorithms, only 100 signals are used for training and classification. Above all, it is interesting to see that the performance of every algorithm becomes quite close with each other as sample size increases.

It was observed through the experiments that, once the overall probability of error is found for a certain predetermined threshold (or the probability of error of one class), the same algorithm can be repeatedly used to produce an asymptotic minimum error probability using new thresholds which are set equal to the overall error probability found from the former calculation. So, if a set of training samples of known classes are given, the threshold which yields minimum error probability for a nonparametric algorithm can be determined. The minimum error occurs when the error probability of one class is the same as the other class if the a priori probabilities of the two classes are the same. This fact is considered more intensively in the Rayleigh distribution case.

Predetermined α -error probabilities of the nonparametric methods and the experimental results are compared in the Table III-5. When the sample sizes are small, the experimental results of error probabilities are not in agreement with the predetermined values. They become closer to predetermined values when the sample sizes increase. For sample size of 16, the differences between the predetermined and the resultant values are in the range of two to three per cent which is also the range for Gaussian distribution case.

The trend of overall error probabilities of an algorithm with respect to the sample sizes is considered in this experiment [Figure 3-6]. The sign test has the highest sensitivity to the changes of sample sizes while the algorithm with the Gaussian assumption has the least range of change. As it was mentioned before, the sign test is very crude in its nature, hence it is very much dependent on the number of samples available to classify. As a whole the signed-rank method works better than linear classifier based on Gaussian assumption and is very competitive with the optimal classifier. The sign test is too crude to use for very small sample size but it is useful for fairly large number of samples. The sign test works almost as good as any other classifier for sample size of 16.

Table III-5. Predetermined and experimental probabilities of error in exponential distribution case

Algorithm		signal level	0.30	0.50	0.75	1.00	1.25
Signed-rank Method	4	Predetermined value	0.350	0.350	0.250	0.250	0.150
		Experimental value	0.418	0.430	0.285	0.315	0.196
	8	Predetermined value	0.390	0.280	0.200	0.120	0.100
		Experimental value	0.423	0.304	0.274	0.167	0.107
	16	Predetermined value	0.350	0.250	0.150	0.070	0.045
		Experimental value	0.333	0.267	0.133	0.077	0.015
Sign Test	4	Predetermined value	0.062	0.062	0.062	0.062	0.000
		Experimental value	0.048	0.055	0.048	0.066	0.000
	8	Predetermined value	0.145	0.145	0.145	0.035	0.035
		Experimental value	0.200	0.200	0.100	0.100	0.100
	16	Predetermined value	0.200	0.200	0.100	0.100	0.100
		Experimental value	0.218	0.233	0.089	0.10	0.089

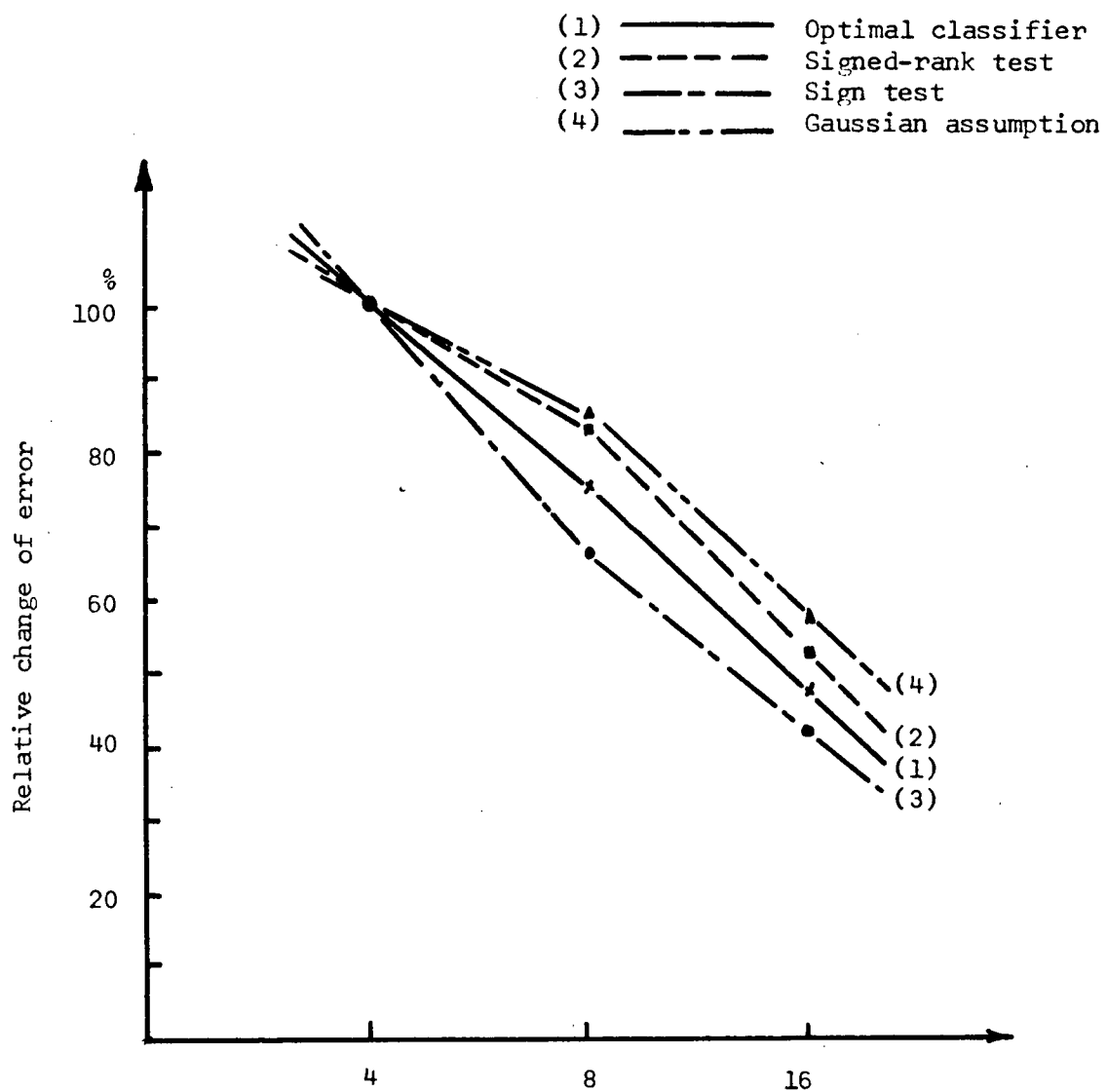


Figure 3-6. Relative changes of error probabilities in accordance with sample sizes

D. Rayleigh Distribution Case

Random variables with Rayleigh distribution are used as the last case for comparison of performance. Different mode values for the two classes of signals are used instead of the different mean values used for exponential and Gaussian distribution cases. Three different mode values of $\sigma_1 = 0.7, 0.8, 0.9$ are used for signal one while a fixed value of mode $\sigma_0 = 1.0$ is used for signal zero. Three different sample sizes are used as before. The sample sizes used are $n = 4, 8$ and 16 . Thresholds for different algorithms are decided as follows.

The Bayes' optimal threshold is determined as

$$\frac{f(\underline{x}/H_0)}{f(\underline{x}/H_1)} = \frac{\prod_{i=1}^n f(x_i/H_0)}{\prod_{i=1}^n f(x_i/H_1)} = 1$$

if the a priori probability of each class occurring is the same as the other and the risk of making a decision is the same for all signals. Independent and identical distributions of samples are assumed.

The resulting classifier decision rule is

$$\sum_{i=1}^n x_i^2 - 4n \left(\frac{\sigma_0^2 \sigma_1^2}{\sigma_0^2 - \sigma_1^2} \right) \ln(\sigma_0/\sigma_1) = 0$$

If the calculation of the above for a given \underline{x} exceeds zero, signal zero is determined. Otherwise, signal one is determined. The threshold for Bayes' decision with the assumption that the distributions are Gaussian is found from the quadratic form of

$$(1/2N_1 - 1/2N_0) \sum x_i^2 + (m_0/N_0 - m_1/N_1) \sum x_i + (n/2)(m_1/N_1 - m_0/N_0)$$

$$- (n/2) \ln(N_1/N_2) = 0$$

where m_i and N_i are the mean and variance of the corresponding Rayleigh distribution. For \underline{x} which makes the above calculation more than zero, signal zero is decided. Otherwise, signal one is assigned. For the Rayleigh distribution with the pdf $f(x) = (x/\sigma^2) \exp(-x^2/2\sigma^2)$, the expected value (mean value) $m = \sigma(\pi/2)^{1/2} = 1.253\sigma$ and the variance $N = \sigma^2(4 - \pi)/2 = 0.4292\sigma^2$, respectively [34].

Nonparametric signed-rank test which was used for Gaussian or exponential distribution case cannot be used without losing efficiency when the pdf is Rayleigh. The reason is that the signed-rank test is based on the assumption of symmetric distribution of the signal 0 such that $f(x_i) = f(-x_i)$. For the Rayleigh distribution, the condition can not be met by a linear transformation. The sign test on the other hand, can still be adopted as before by shifting the pdf to satisfy the condition $F(0) = 1/2$.

Nonlinear ranking for the signed-rank method may be adopted for this circumstance. Instead of the usual ranking procedures a transformation of data is used to result in a symmetric or near symmetric distribution. However, the transformation of data requires complete distribution information which is not appropriate in the use of nonparametric methods.

Considering the difficulties of using signed-rank method in this experiment, two-input case sign and signed-rank methods are also used by generating independent noise channel data. Results of the experiments

which include the Bayes' optimal classifier, Bayes' classifier with Gaussian assumption, signed-rank test, sign test and K-class algorithm are given in Figures 3-7, a, b, c, and the experiments for the two-input channel sign and signed-rank test are compared in Figure 3-8.

The signed-rank test does not perform as well as the optimal test in this distribution case. The reason is as stated in page 54. Sign test which is already known to be too crude for small numbers of samples displayed itself again as a poor classifier. For the sample sizes four and eight, it resulted in error probabilities which are too large for practical use compared to other classifiers. Sign and signed-rank test applied for two-input case also give large error probabilities compared to the optimal classifier. The classifier based on the assumption of Gaussian distribution works very good over the entire range of experimental conditions. There is very little advantage to use the optimal classifier instead of adopting the Gaussian assumption since there is less than one percent of error probability difference on the average by using the optimal classifier. The reason for this extraordinary performance of the classifier based on the Gaussian assumption seems to be that the Rayleigh distribution becomes similar to the Gaussian as the mode value increases.

As in the other experiments already seen, the K-class algorithm works good in most of the varied circumstances. This algorithm seems a little inferior to the optimal classifier for relatively large signal separation (mode value difference) but it works better than any other algorithm for small signal separations. Irregular change of error probabilities for different data conditions like sample sizes and

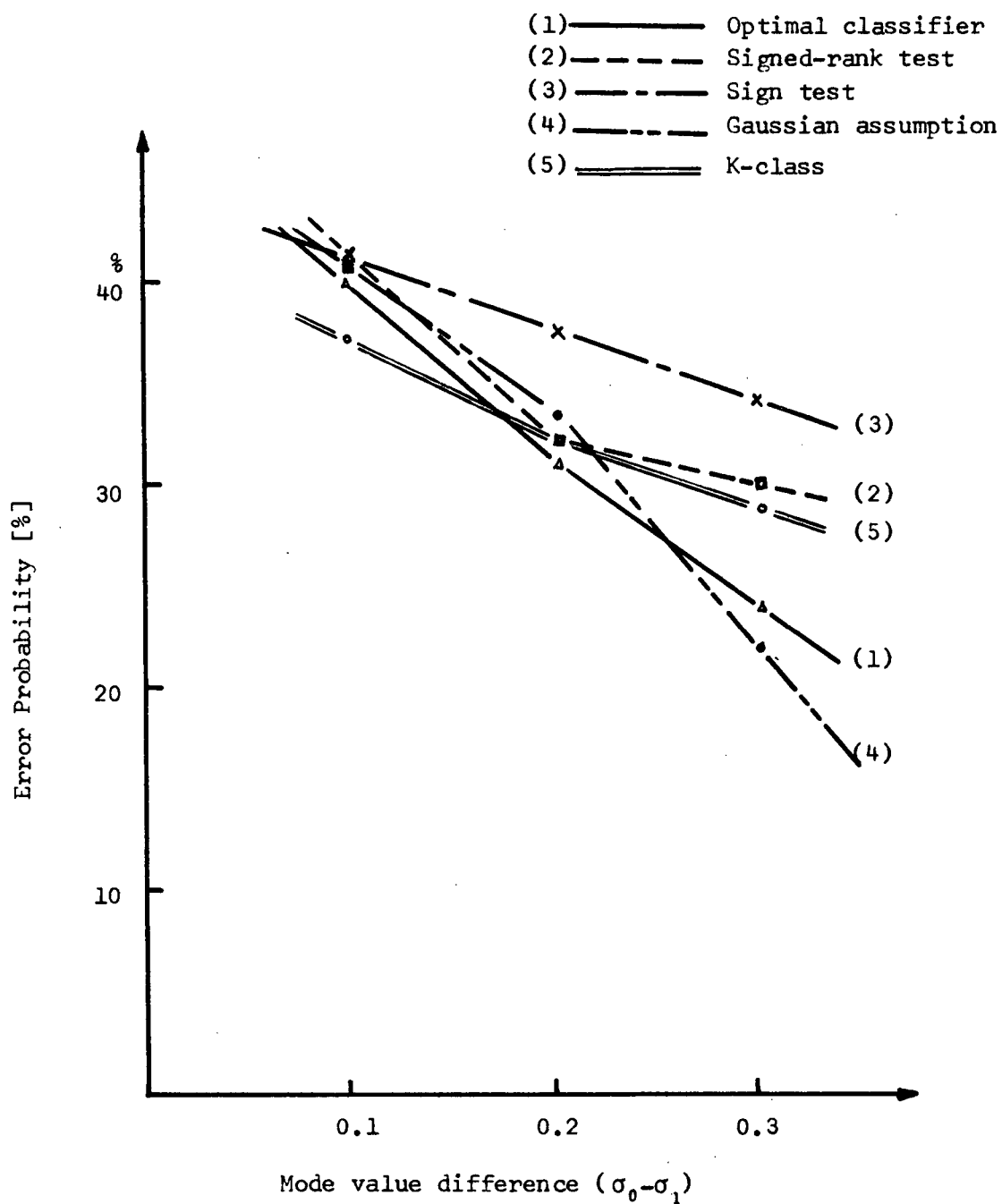


Figure 3-7.a. Error probabilities of different algorithms with Rayleigh distribution (4 features)

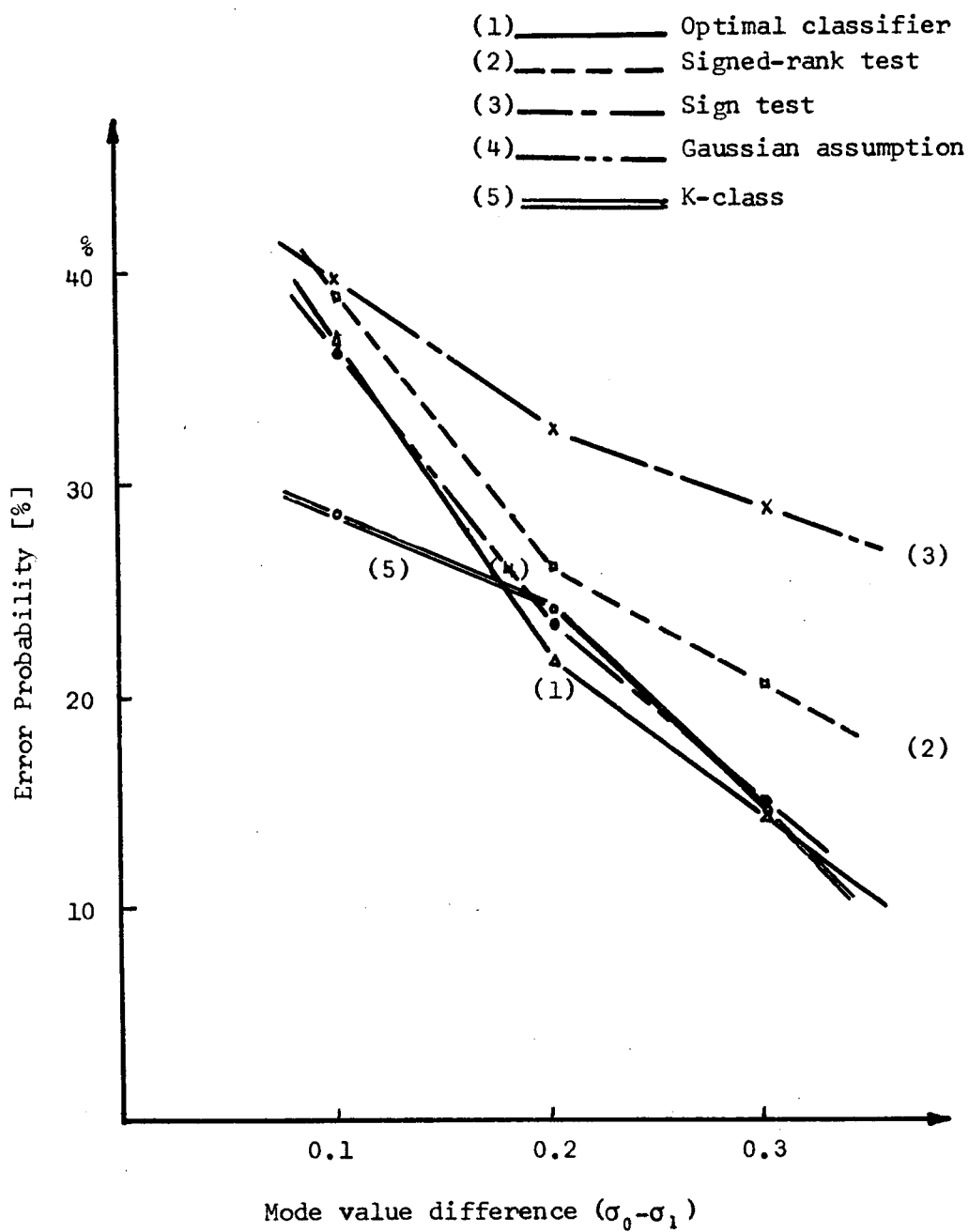


Figure 3-7.b. Error probabilities of different algorithms with Rayleigh distribution (8 features)

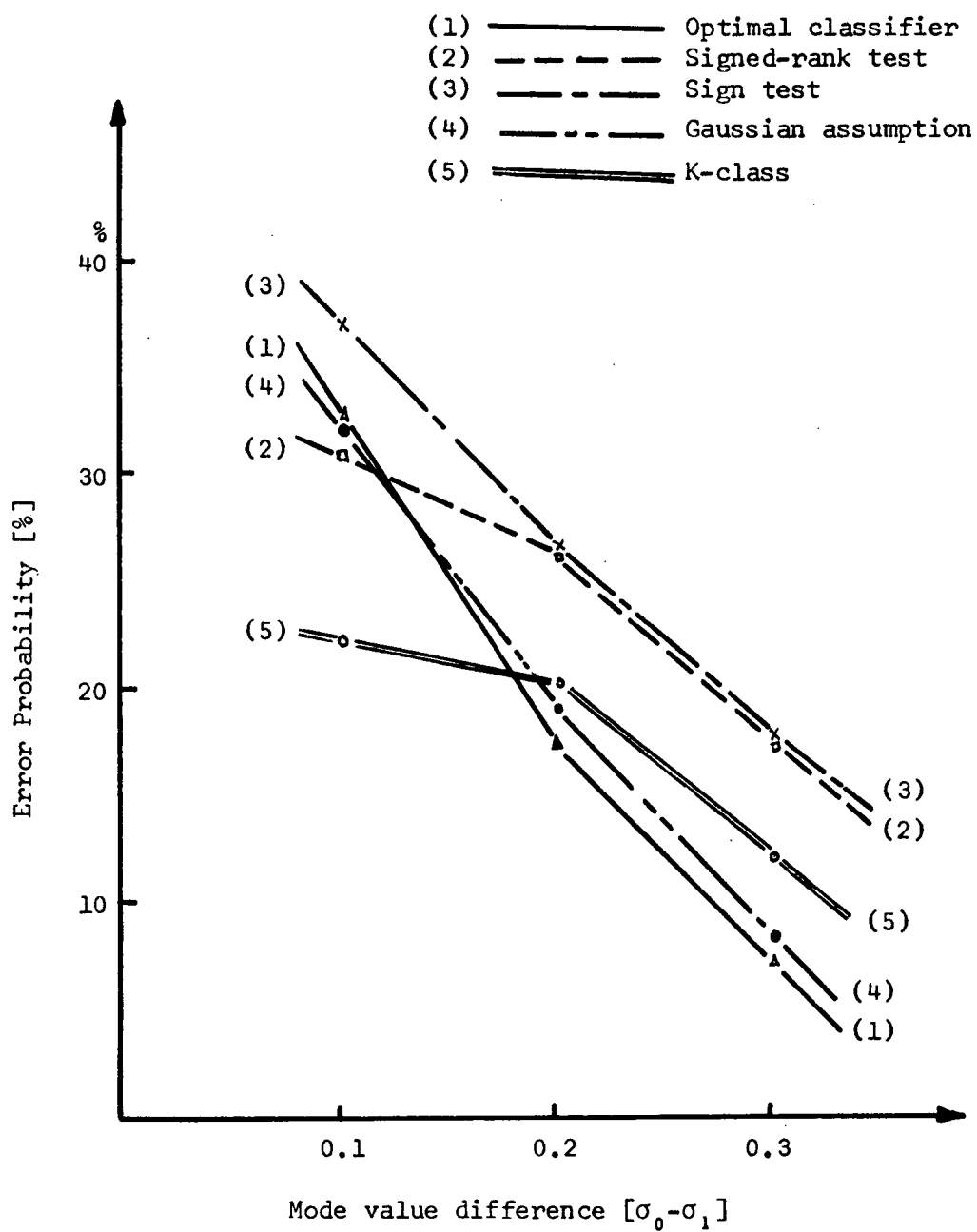


Figure 3-7.c. Error probabilities of different algorithms with Rayleigh distribution (16 features)

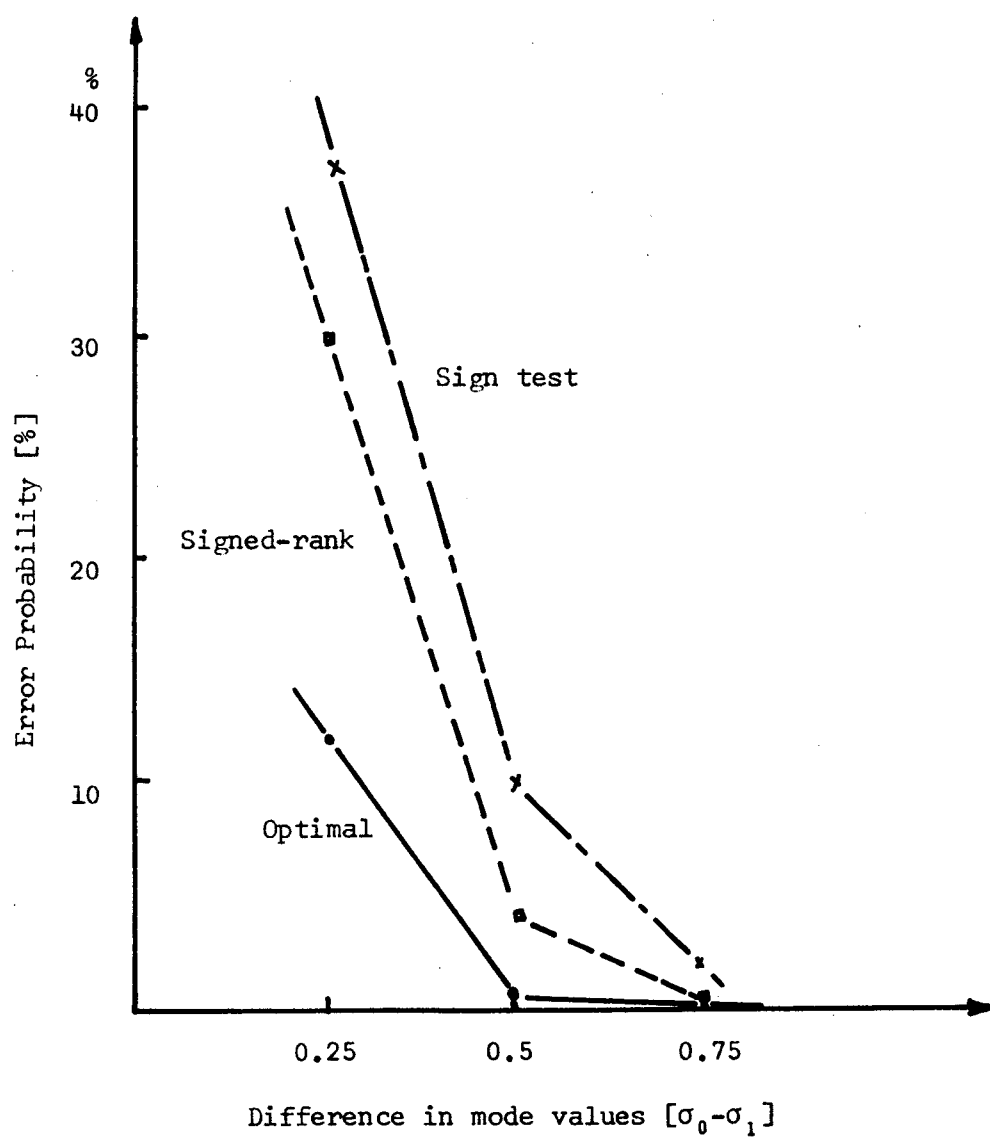


Figure 3-8. Two-input nonparametric tests of Rayleigh distribution.

signal-to-noise ratios is observed again as in the previous experiments.

One of the reasons for this irregularity is that fewer signals are used in this K-class algorithm than in the other algorithms.

In general, nonparametric methods seem to be inferior to parametric methods for the Rayleigh distribution. Only for large sample sizes, say $n = 16$, and small signal-to-noise ratio their usefulness predominates.

One concept is worth noting. Nonparametric methods seem to be less sensitive to the sample sizes and signal level differences. The relative changes in error probabilities of Bayes' optimal classifier and the nonparametric methods for different mode values are seen in Figure 3-9. The relative changes of error probabilities on the average for different mode values have the least slope for sign test while the algorithm with Gaussian assumption has the steepest slope of all. Though it does not necessarily imply the usefulness of nonparametric tests, the robustness does show that sign test or signed-rank test is viable for an algorithm with other small signal-to-noise ratio situations. It was emphasized several times before that the overall probability of error of a nonparametric algorithm is dependent on the predetermined error probability of one class. For the Rayleigh distribution, tests were run to observe the actual behavior of the error probabilities according to the changes of threshold values which determine α -error probabilities. Two mode values $\sigma_1 = 0.7$ and 0.9 are assigned to the distributions of class one while class zero

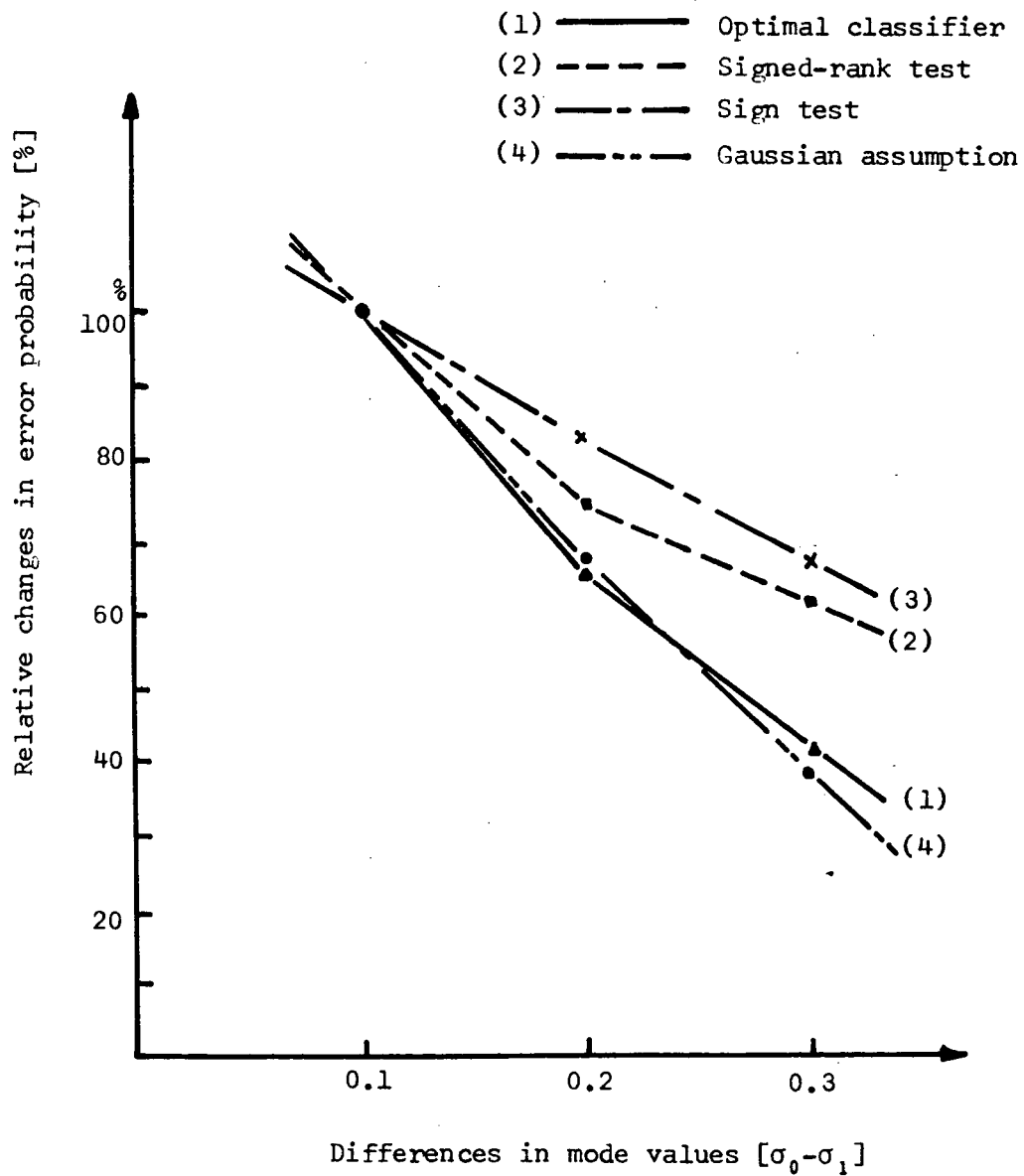


Figure 3-9. Relative changes in error probabilities for different mode values

has a fixed mode value of $\sigma_0=1.0$. A sample size of 16 is used in both signed-rank and sign tests. Results are shown in Figures 3-10.a, b, c. In the figures it is noticed that the minimum overall error probabilities occur at or near the thresholds at which both α and β -error probabilities become equal. This phenomenon is more apparent when the difference of mode values of both classes is larger. The discrepancy of having a minimum error probability at a threshold value other than that which makes the α and β -error equal in the Figure 3-10.d may be eliminated by using a larger number of samples.

E. Complexity of Calculation of Each Algorithm

The complexity of calculation for the specified algorithm is one of the most important factors in the practical application. Each algorithm has a unique process of data treatment. It is compared to other algorithms for its requirements on calculations in this section.

A nonparametric sign test needs only n comparisons of signs and n summing operations on integer numbers for n input data. It also needs only a couple of memory cells for a threshold and a summed integer number of signs. This is the least complex algorithm of all.

The signed rank method should rank the absolute values of n observed data and take the sum of ranks of positive observations, hence, it requires n operations of taking absolute values, $n(n+1)/2$ comparison steps for ranking and n comparisons of signs and n summing operations. This requires at least $2n$ plus a few memory cells. Apparently signed-rank method takes much more time for data processing than the sign test and a linear classifier do but it has no multiplication or division

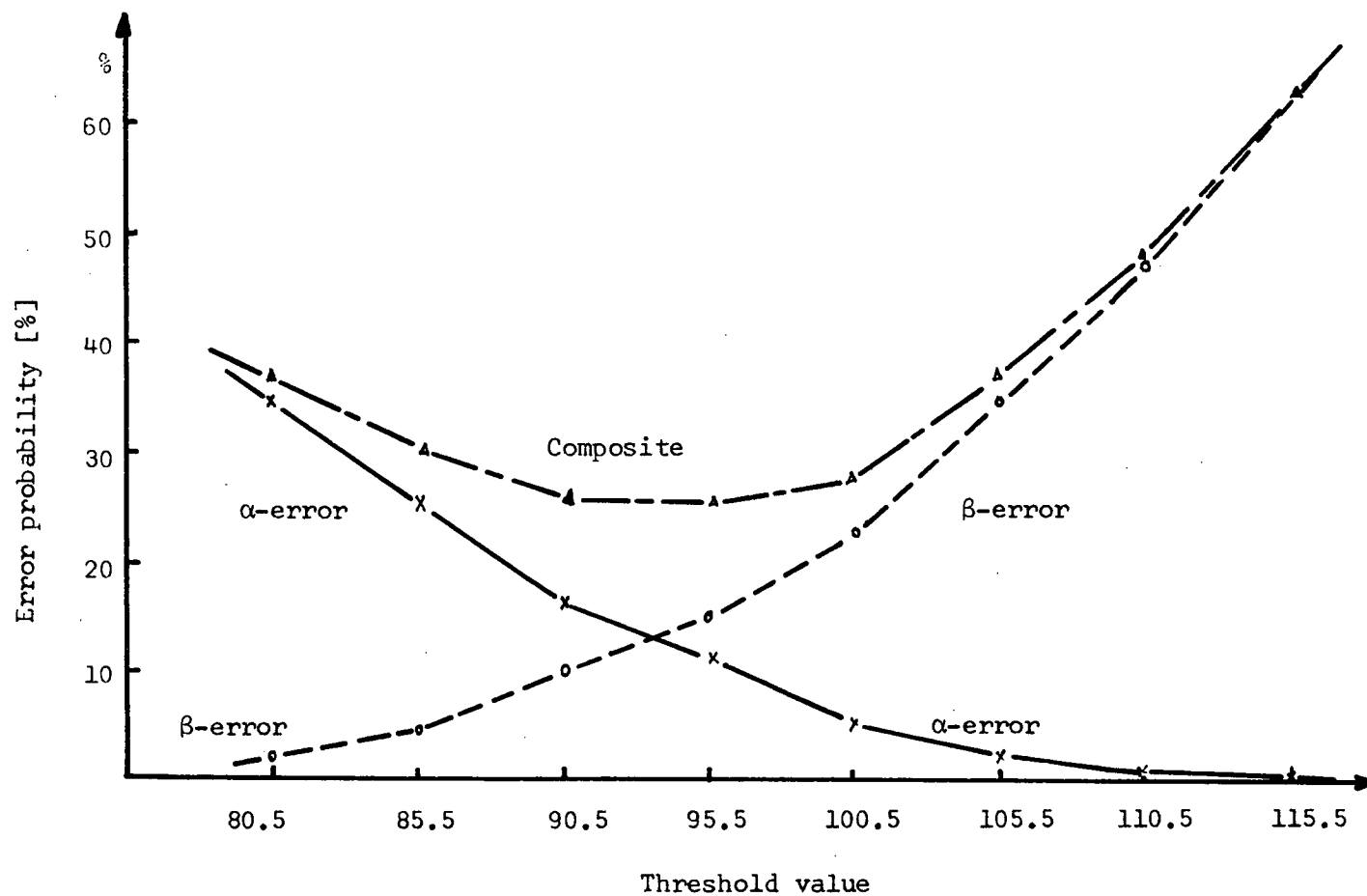


Figure 3-10.a. Changes of α -and β -error probabilities according to thresholds in signed-rank test. (Difference in mode values = 0.3. Rayleigh distribution)

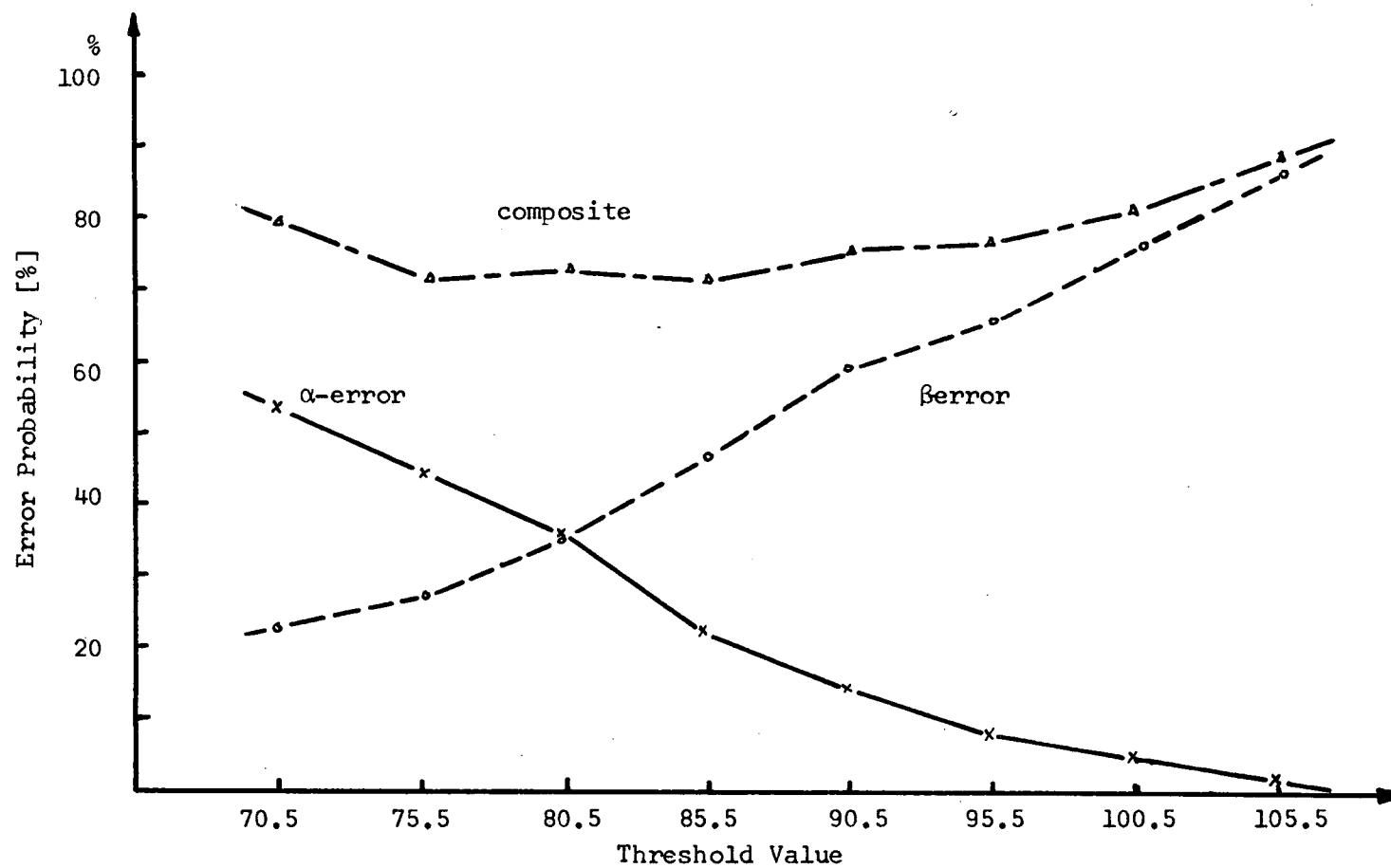


Figure 3-10.b. Changes of α - and β -error probabilities according to thresholds in signed-rank test (Difference in mode values = 0.1. Rayleigh distribution)

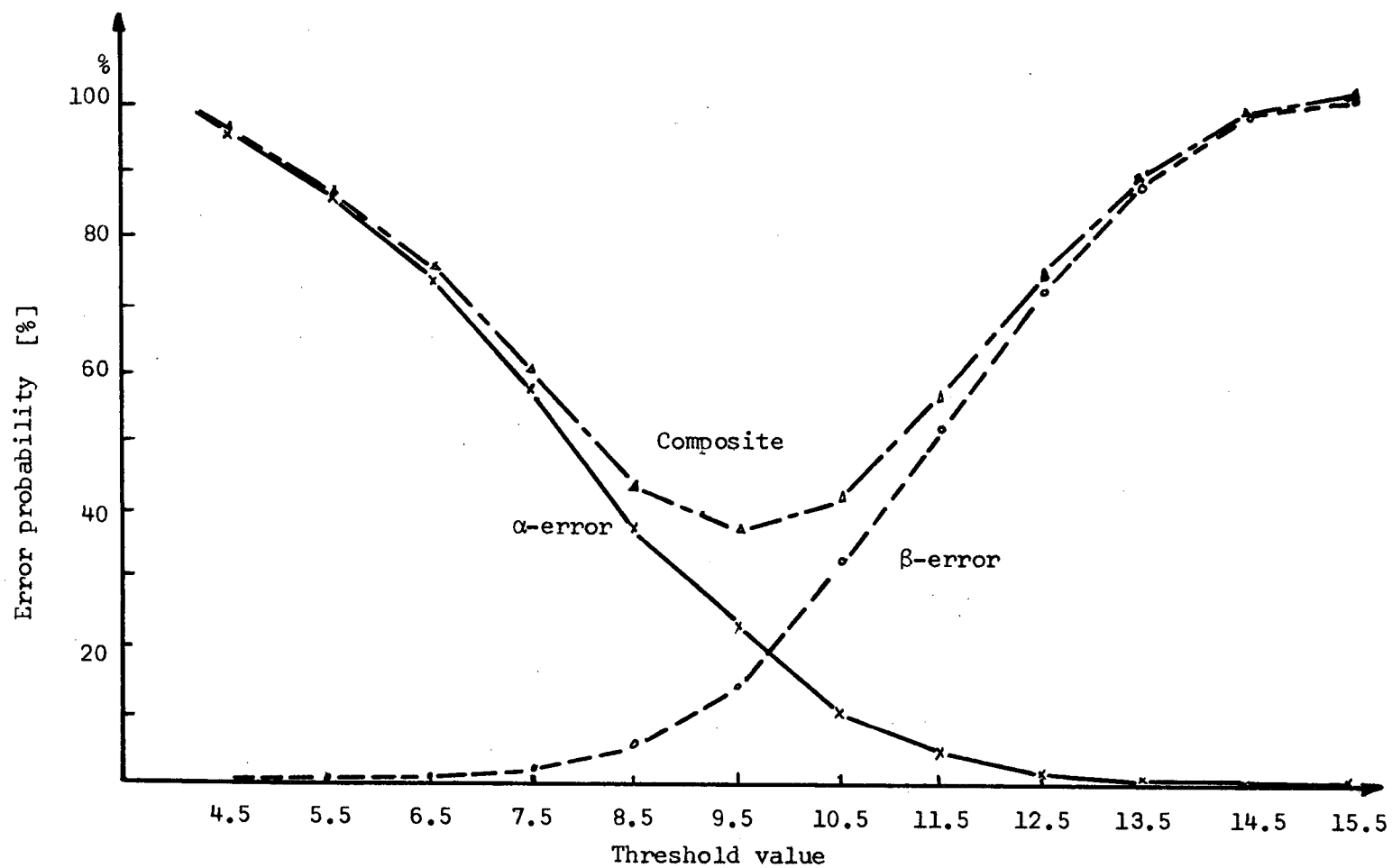


Figure 3-10.C. Changes of α - and β -error probabilities according to thresholds in sign test
(Difference in mode values = 0.3. Rayleigh distribution)

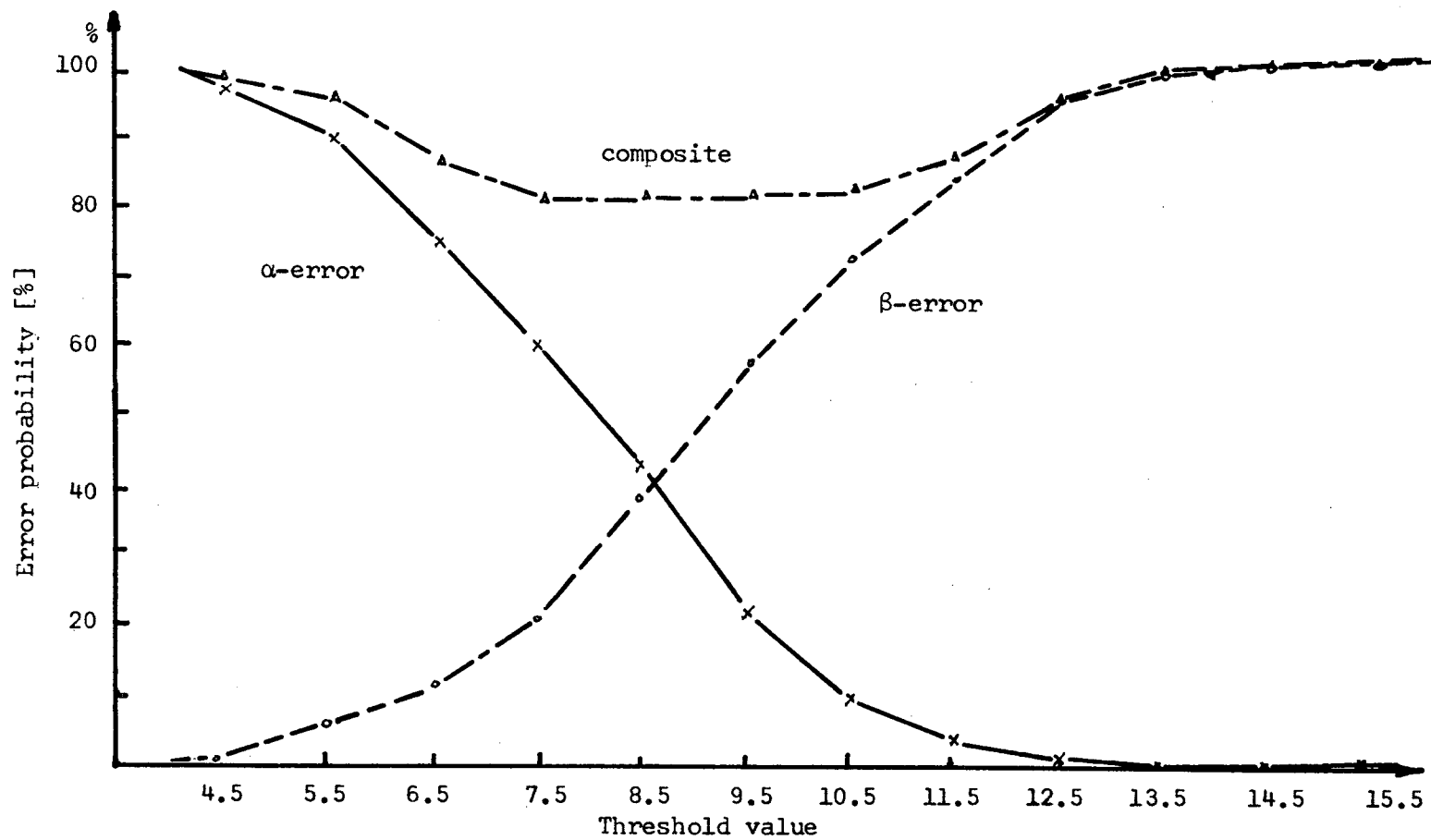


Figure 3-10.d. Changes of α - and β -error probabilities according to thresholds in sign test
(Difference in mode values = 0.1. Rayleigh distribution)

operations which are present in some other classifiers.

A linear classifier which is based on independent and identical Gaussian noise needs essentially n summing and one thresholding operation to make a decision on n observed data. Only a couple of memory cells are necessary. Of course, the variances of both classes are assumed same. If the variances differ from each other, the optimal classifier should perform n multiplication and $2n$ summations in addition to a few thresholding operations. The multiplication takes much more time than adding, subtracting or comparing a set of data. For the data with two-sided exponential distribution, $2n$ summations, n subtractions and $2n$ absolute values are necessary to make a decision. A few memory cells are required. If the data is Rayleigh distributed, n summations and n multiplications on input data and one threshold operation is necessary for the optimal decision. The memory storage required is small. The above is summarized in Table III-6.

Table III-6. Calculations involved in each algorithm

Algorithms		Addi- tion	Subtrac- tion	Compari- son	Multipli- cation	Absolute value	Memory required
Sign test		n	0	n	0	0	less than 3
Signed-rank		0	0	$n + \frac{n(n+1)}{2}$	0	n	$2n + \text{a few}$
O P T I M A L	Gaussian equal variance	n	0	0	0	0	less than 3
	Gaussian different variance	$2n$	0	0	n	0	less than 10
	Laplacian	$2n$	n	0	0	$2n$	less than 5
	Rayleigh	n	0	0	n	0	less than 3

F. Summary of the Chapter

The performances of nonparametric classifiers and Bayes' parametric optimal classifier are compared. The Gaussian assumption and the K-class algorithms are also used for additional comparisons. These methods are applied to the noise distributions of the form of Gaussian, two-sided exponential or Laplacian and Rayleigh.

Nonparametric methods work very well for Gaussian and Laplacian distribution cases. Even the sign test has more efficiency than the linear classifier for relatively large samples like $n = \text{eight}$ and more when the distribution is Laplacian. However, these nonparametric tests give considerably larger error probabilities for the Rayleigh distribution case, where the distributions are not symmetrical. This requirement of symmetric distribution seems to be the major disadvantage of the nonparametric signed-rank test.

Two-input nonparametric methods generally failed. Even more, it requires two independent input channels which are not easy to find in a practical situation.

As a whole, the nonparametric sign test looks attractive as an algorithm when the signal-to-noise ratio is very small and there are enough samples. Sample sizes of more than four are needed for satisfactory results. The signed-rank method is also very useful for the symmetrical distributions. However, the calculation complexity of this test increases rapidly as sample size increases and is not favorable compared to the linear classifier. The linear classifier which is based on the Gaussian distribution assumption works well for most of

the experiments producing consistent results which are comparable to optimal classifier's.

The nonparametric threshold which gives an asymptotic minimum error probability can be found by repeated adjustment of thresholds if a set of sample vectors of known classes is given.

The ARE of a nonparametric method may not be a general performance index since the actual efficiency of one method compared to the other is changing because of the different signal-to-noise ratio and the number of samples. But it still gives a very good idea of the relative performance of the algorithms.

The limitations using the nonparametric methods are the requirements on the data distributions such as: the statistical independence between each data, the identical distribution of each other and the continuous and symmetrical distribution of the variables. Symmetric condition is required especially for the signed-rank method.

The multivariate, multi-class problem is considered in Chapter IV.

CHAPTER IV

APPLICATIONS TO THE MULTI-CLASS PROBLEMS

Throughout previous chapters only the univariate, two-class problems are considered. In the practical pattern recognition problems, however, the general nature of input data are multivariate and the decision-making is usually multi-class conditioned. The generalization of the two-class problem into a multi-class, multivariate problem is considered in this chapter.

Since the nonparametric methods already discussed in univariate cases have inherent limitations like independent sampling of data and symmetrical distributions for each class in case of the signed-rank test, there must be modifications of the nonparametric methods to apply the methods to multi-class and multivariate situations.

For clarity of understanding the problems, the multi-class, multivariate problems are grouped into several categories according to the nature of the variable: (1) univariate, multi-class case, (2) multivariate, two-class case, (3) multivariate, multi-class case. They are discussed in the following sections.

A. Univariate, Multi-class Problems

Since the nonparametric method essentially tests a composite hypothesis, i.e., it merely tests whether the null hypothesis is true or not, this method needs at most k independent statistical tests for k different classes. If the data are from univariate distribution, the methods used in the two-class problems can be applied in a repetitive way to the multi-class problems.

1. Sign test

This test is applicable when the conditions required in the two-class problems are satisfied. They are the continuity of distributions over the range and the differences in the median values of the k different classes. The data obtained should be independent of each other. Let $\underline{m}_1, \dots, \underline{m}_k$ be the median vectors of each of k classes, and \underline{x} be the measurement vector with n observations. The two-class sign test is applied for each pair of $\underline{x} - \underline{m}_j$; $j=1, 2, \dots, k$. If the \underline{x} has been from the j -th class, the number of the positive and negative signs would be almost equal for the data $\underline{x} - \underline{m}_j$. For the rest of the classes the value $\underline{x} - \underline{m}_i$; $i \neq j$, would show a larger number of positive or negative signs than the number of opposite signs. So, after determining positive and negative signs of each of k different data sets, $\underline{x} - \underline{m}_i$; $i=1, \dots, k$, the vector \underline{x} is assigned to the class at which the difference of numbers between positive and negative signs is the minimum. Naturally occasions when there are more than one class which yield the same minimum difference in numbers of positive and negative signs may happen especially for small number of observations. There seems to be no way out of this confusion. Hence, a sufficiently large number of observations is necessary for this test.

2. Signed-rank test

This test is also the direct generalization of the two-class signed-rank test, and is sensitive to the differences in mean values between the classes of symmetrical distributions. Let there be k classes as before and let \underline{x} be the measurement vector of n observations. Assume $\underline{\mu}_1, \underline{\mu}_2, \dots, \underline{\mu}_k$ to be the mean vectors of the k different classes.

The test then follows the procedures described in two-class problem for each of $\underline{x} - \underline{\mu}_i$; $i=1, \dots, k$, data sets. For example, at the j -th test, find the difference of $\underline{x} - \underline{\mu}_j$ and find the rank of each element in increasing order of absolute magnitude of the difference. If the sample \underline{x} is from the j -th class, the fundamental conditions of independent samples are insuring that the sum of ranks from positive differences will be about equal to the sum of ranks from negative differences. So, after determining positive and negative signed-rank sums of each of k tests, the data set \underline{x} is assigned to the class for which the rank sum of negative differences is closest to that of positive differences.

The test may be terminated before k steps are taken. During the test, the data set \underline{x} may be assigned to the class at which the signed-rank sum is within a certain significance level, which can be determined through the same way used in a two-class problem.

3. Rank sum test

Compared to the sign test, the signed-rank test is much more efficient as it was seen in two-class problems but it imposes a serious restriction which is that the data distributions are symmetrical. It is thus necessary to adopt an algorithm which is more general than those discussed. The rank sum test is used instead for testing the differences in mean values of different classes whose distributions need not be symmetrical but identical in shapes for all classes.

As in the two-class problems, rank sum test can be used for relatively general hypothesis testing but it needs additional independent data sets which represent k different classes. The two-class rank sum tests are executed in turn to the k paired sets of

data; \underline{x} and the sample vector of each class. By the same reason stated in the two-class problem, the rank sum of the observed vector \underline{x} will be distributed in a statistically fixed form for the null hypothesis that the data are from the j -th class. Instead of only one null hypothesis of the two-class problem, there are k independent null hypotheses for the multi-class problem.

Using the fixed distribution function determined from the null hypotheses, all the k rank sums are checked correspondingly to get the probabilities of these sums occurring. The data set \underline{x} is assigned to the j -th class if the probability of the j -th rank sum is the largest.

B. Multivariate, Two-class Problems

Before proceeding to the multivariate multi-class problem, it seems necessary to consider the multivariate two-class problems to see the nature of the multivariate case. Let $\underline{x} = \{x_1, \dots, x_n\}$ be the observation vector from one channel and $\underline{y} = \{y_1, \dots, y_n\}$ from another. The multivariate two-class rank sum test can be applied for this case. In the previous example of univariate data $\underline{x}, \underline{y}$, where all of the x_i 's and y_i 's are identical and independent, the nonparametric rank sum method makes use of the ranks of the combined data to test the null hypothesis that the two data sets are from the same distribution against the alternative that they are not. The test essentially is based on the numbers M_0, M_1, \dots, M_n where M_i is the number of y 's falling between the i -th and $(i+1)$ st ordered x 's. When the observations x 's and y 's are from multivariate distributions then the number M_i which gives precise statistical equivalence to the univariate situation is not readily decided. First there must be determined the hyperplane blocks

[23] which are the multivariate analogy to the univariate regions between the ordered x_i 's. A deeper study in determining blocks is called for but it is not tried in this work. Once the equivalent blocks are found, the procedures of handling the rank data remain to be the same as those of univariate two-class problems.

C. Multivariate, Multi-class Problems

Most of the multi-class problems discussed in some publications [1], [23] test the null hypothesis that all of the k sets of data are from the same distribution against the alternative that there is significantly different distribution in data. Since this hypothesis testing is not sufficient for identifying each of k -classes, a different algorithm must be developed.

One possible way to treat this problem seems to be to apply the multivariate, two-class algorithms to \underline{x} , the observed data, k times with k different sets of samples, each sample representing the typical distribution of one of k classes. Eventually k different probabilities which are the probabilities of \underline{x} being from each of the k classes will be obtained. \underline{x} is then assigned to the class for which the probability obtained is the highest.

This multivariate, multi-class problem is very difficult to treat and the above suggestion must be proven in practical circumstances.

D. Summary of the Chapter

A univariate multi-class problem is mainly considered in this chapter. Repetitive applications of two-class algorithms accomplish the job. If the problem is multivariate, the transformation of multivariate data to univariate data is necessary. Finding the blocks,

which are statistically equivalent to the regions bounded by ordered x 's in univariate case is the main problem. This area needs more study.

The most general case, multivariate multi-class problem, might be solved by repetitive use of multivariate two-class algorithm, but no attempt is made to simulate the problem since this gets too involved and the merit of nonparametric methods will be lost in the complexity of calculations. Many problems may be solved more easily and practically assuming univariate situations.

CHAPTER V

CONCLUSIONS

A. Summary

The nonparametric methods were compared to the optimal parametric classifiers and the K-class algorithm. The nonparametric methods performed very competitively for most of the conditions subjected with some exceptions. They worked especially good when the sample sizes were large.

The signed-rank test was almost as good as, and sometimes better than the optimal classifier but the test needs somewhat higher complexity of calculations compared to parametric tests for the density functions studied. This disadvantage may be excused when the distributions are not simply Gaussian, Laplacian or Rayleigh's where ordinary optimal classifiers need simple calculation steps. Nonparametric methods, however, have fixed procedures that do not vary with the distribution shape. Another significant drawback of the signed-rank test is its requirement of symmetric data distribution of each class. Since this requirement is hard to be satisfied in practice, symmetric conditions may be assumed at the expense of the efficiency of the test as it was done in the Rayleigh distribution case.

When the sample size is large and fast data processing is necessary, the sign test is a very useful method. This sign test needs only a few simple integer arithmetic operations for data processing and its efficiency is good for most of the distributions. A mixed statistical test which employs both signed-rank and sign test looks attractive as the simplicity of sign test is combined with the efficiency of signed-

rank test and the compromise between the two is made.

Tests with two-input channels seemed to be too inefficient for practical use. The rank sum test is identical to the signed-rank test if the distributions of the variables are symmetrical. This test is sensitive to the differences of medians in two identically or symmetrically distributed data. One major demerit of this test is that it requires independent input channels. The above were observed through the results of simulations by computer and were depicted in figures of Chapter III.

The ARE does not give a direct efficiency of an algorithm for different sample sizes and signal-to-noise ratios, but it still shows the relative figure of merit at large of one classifier to another.

The optimal nonparametric thresholds were determined by taking those for which the α - and β -error probabilities of the two classes are the same. This phenomenon was also experimentally seen in Chapter III.

It was observed that the K-class algorithm competed very well among other algorithms but the distributions had to be unimodal to be efficient in classification.

The generalization of the univariate, two-class problems into the multivariate, multi-class problem was considered. The univariate, multi-class problem was solved by repeated applications of the univariate, two-class algorithms. For the most general case, the multivariate, multi-class problems, no specific conclusion was able

to be drawn.

B. Suggestions for Further Study

There are five most imminent areas of research to be done. First of all, the applications of the nonparametric methods to the real data obtained from the photographic imagery are desired to verify the practical usefulness of the methods. A univariate, multi-class algorithm may be used with reasonable assumptions. Nonlinear ranking techniques, in the case of nonsymmetric distribution, need to be investigated further as the second research area. This technique is necessary to employ the high efficiency of the signed-rank test for nonsymmetric data distributions.

The third research area includes the determination of optimal threshold for nonparametric methods when the a priori probabilities of the two classes are different. The efficiency of the K-class algorithm using the data which are not used to train the algorithm should be investigated for more direct comparisons with other methods. The last research area is to investigate more on the multivariate, multi-class problems. The determination of blocks, which is in analogy to the regions of ordered univariate data, should be studied.

GLOSSARY OF TERMS

x	Random variable.
\underline{x}	A random vector with n elements. The underline specifies a column vector.
$X(m)$	A set of m random vectors \underline{x} . A (m,n) matrix is implied.
$f(x)$	Probability density function (pdf) of a random variable x .
$F(x)$	Cumulative distribution function (cdf) of a random variable x .
y, \underline{y}	Same as x, \underline{x} except that these are input from different channels.
μ_i	Expected value (or mean value) of a random variable. Subscript represents i -th class.
σ_i^2	Variance or a central moment of a random variable of i -th class.
Σ	Variance and co-variance matrix for multi-class case.
\bar{x}	Sample mean.
s^2	Sample variance.
H_0	Null hypothesis that noise only is present in the input channel of a classifier.
H_1	Alternative hypothesis. Signal is assumed to be present in the input.
α	The error of the first kind, or the probability of misclassifying a set of data as class 1 while the data are actually from class 0. Equivalent to the probability of falsely rejecting H_0 .
β	The error probability of the second kind, or the probability of misclassifying a set of data into class 0 while the data

are actually from class 1. Equivalent to the probability of falsely rejecting H_1 .

$p(i)$	A priori probability of class i .
$L(x)$	Likelihood ratio.
K_i	Cost of making a decision of the i -th class.
$\text{erf}(x)$	Error function of x .
C	Threshold for a classifying algorithm.
r	A rank of an observation x among the set of absolute x_i 's in increasing order.
\underline{r}	A vector composed of r .
$g(x)$	Decision function (discriminant function)
$p(x)$	Probability density function of x . This is the same expression as $f(x)$, but $p(x)$ is used mainly in the parametric case.
$e_{1,2}$	A relative efficiency of a method 2 compared to another method 1.
$\text{ARE}_{1,2}$	An asymptotic relative efficiency of a method 2 compared to another method 1.
\rightarrow	The arrow is used for either one of the words or the set of words: implies, is concluded as, or if...then .

APPENDIX A

REDUCTION OF A QUADRATIC FORM TO A LINEAR FORM

For the multivariate Gaussian noise which is added to the dc signal, the quadratic form which is the logarithm of the likelihood ratio can be reduced to a linear form by applying summation calculation, without knowing the characteristics of quadratic form. This, of course, is possible when the distribution functions have the same covariances $\sum_0 = \sum_1 = \sum$.

By definition,

$$(\underline{x} - \underline{\mu}_0)^T \sum^{-1} (\underline{x} - \underline{\mu}_0) = \sum_i^n \sum_j^n (x_i - \mu_{0i})(x_j - \mu_{0j}) \sigma_{ij}$$

and

$$(\underline{x} - \underline{\mu}_1)^T \sum^{-1} (\underline{x} - \underline{\mu}_1) = \sum_i^n \sum_j^n (x_i - \mu_{1i})(x_j - \mu_{1j}) \sigma_{ij}$$

Then,

$$\begin{aligned} \ln L(\underline{x}) &= -\frac{1}{2} \sum_i^n \sum_j^n [(x_i - \mu_{0i})(x_j - \mu_{0j}) \sigma_{ij} \\ &\quad - (x_i - \mu_{1i})(x_j - \mu_{1j}) \sigma_{ij}] \\ &= -\frac{1}{2} \sum_i^n \sum_j^n (x_i x_j - x_i \mu_{0j} - x_j \mu_{0i} + \mu_{0i} \mu_{0j} - x_i x_j \\ &\quad + x_i \mu_{1j} + x_j \mu_{1i} - \mu_{1i} \mu_{1j}) \sigma_{ij} \\ &= -\frac{1}{2} \sum_i^n \sum_j^n [(\mu_{1j} - \mu_{0j}) x_i + (\mu_{1i} - \mu_{0i}) x_j] \end{aligned}$$

$$\begin{aligned}
& +\mu_{0i}\mu_{0j}-\mu_{1i}\mu_{1j}]\sigma_{ij} \\
& = \sum_i^n \sum_j^n x_i(\mu_{0j}-\mu_{1j})\sigma_{ij} - \frac{1}{2} \sum_i^n \sum_j^n (\mu_{0i}\mu_{0j}-\mu_{1i}\mu_{1j})\sigma_{ij}
\end{aligned}$$

Since the last term of the above equation is a constant for any \underline{x} ,

$\ln L(\underline{x}) = \underline{x}^T \underline{\Sigma}^{-1}(\underline{\mu}_0 - \underline{\mu}_1) + \text{constant}$ which is a linear polyromial. The constant value is sometimes called the bias.

APPENDIX B

ESTIMATION OF PARAMETERS USING REPETITIVE CALCULATIONS

The likelihood ratio test of the two distributions without any parameter value given but only with the sample vectors of known classes encounters the problem of estimating the parameters by the use of given sample vectors. Then the likelihood ratio is

$$L(\underline{x}) = \frac{f(\underline{x}/X^0(m), H_0)}{f(\underline{x}/X^1(m), H_1)}$$

where $X^i(m) = \{\underline{x}^i(1), \dots, \underline{x}^i(m)\}$, $i=0,1$ which is the set of m sample vectors of class i . However, the numerator and denominator can be written as

$$f[\underline{x}/X^i(m), H_i] = \int_{-\infty}^{\infty} f(\underline{x}/\theta, H_i) f[\theta/X^i(m), H_i] d\theta$$

The determination of $f[\theta/X^i(m), H_i]$ is the main problem which is solved in a repetitive way, shown below.

From the Bayes' theorem (not Bayes' criterion)

$$f[\theta/X(m)] = \frac{f(X(m)/\theta) f(\theta)}{\int_{-\infty}^{\infty} f[X(m)/\theta] f(\theta) d\theta} \quad \text{where}$$

the condition H_i and superscript i of X are omitted for convenience.

$$\text{But } f[X(m)/\theta] = \frac{f[X(m) \cdot \theta]}{f(\theta)}$$

$$= \frac{1}{f(\theta)} [f(\underline{x}(1), \underline{x}(2), \dots, \underline{x}(m), \theta)]$$

$$= \frac{1}{f(\theta)} f[\underline{x}(m)/\underline{x}(1), \dots, \underline{x}(m-1), \theta] f[\theta, \underline{x}(1), \dots, \underline{x}(m-1)]$$

$$= \frac{1}{f(\theta)} f[\underline{x}(m)/X(m-1), \theta] f[\theta/X(m-1)] f[X(m-1)]$$

where $X(m-1) = \{\underline{x}(1), \dots, \underline{x}(m-1)\}$

Hence,

$$\begin{aligned} f[\theta/X(m)] &= \frac{\frac{1}{f(\theta)} f[\underline{x}(m)/X(m-1), \theta] f[\theta/X(m-1)] f[X(m-1)]}{\int_{-\infty}^{\infty} \frac{f(\theta)}{f(\theta)} f[\underline{x}(m)/X(m-1), \theta] f[\theta/X(m-1)] f[X(m-1)] d\theta} \\ &= \frac{f[\underline{x}(m)/X(m-1), \theta] f[\theta/X(m-1)] f[X(m-1)]}{f[X(m-1)] \int_{-\infty}^{\infty} f[\underline{x}(m)/X(m-1), \theta] f[\theta/X(m-1)] d\theta} \\ &= \frac{f[\underline{x}(m)/X(m-1), \theta] f[\theta/X(m-1)]}{\int_{-\infty}^{\infty} f[\underline{x}(m)/X(m-1), \theta] f[\theta/X(m-1)] d\theta} \\ &= \frac{f(\underline{x}(m)/\theta) f[\theta/X(m-1)]}{\int_{-\infty}^{\infty} f[\underline{x}(m)/\theta] f[\theta/X(m-1)] d\theta} \end{aligned}$$

Here

$f[\underline{x}(m)/X(m-1), \theta]$ is put equal to $f[\underline{x}(m)/\theta]$ because of conditional independence.

The last term is the desired repetitive form to be used for the calculation of $f[\theta/X^i(m), H_1]$ of the likelihood ratio test.

BIBLIOGRAPHY

A. On Nonparametric Algorithms

- (1) Bradley, J. V., Distribution-free Statistical Tests, Englewood Cliffs, N.J.: Prentice-Hall, Inc., pp.96-114, 129-134, 1968.
- (2) Carlyle, J. W. and J. B. Thomas, "On Nonparametric Signal Detectors", IEEE Trans. on Information Theory, Vol. IT-10, No. 2, pp.146-152, April 1964.
- (3) Chadwick, H. D. and L. Kurz, "Two Sequential Nonparametric Detection Procedures", Report TR 400-155, Department of Electrical Engineering, New York University, New York: March 1967.
- (4) Daly, R. F. and C. K. Rushforth, "Nonparametric Detection of a Signal of Known Form in Additive Noise", IEEE Trans. on Information Theory, Vol IT-11, No.1, pp.70-76, January 1965.
- (5) Feustel, E. A. and L. D. Davisson, "The Asumptotic Relative Efficiency of Mixed Statistical Tests", IEEE Trans. on Information Theory, Vol. IT-13, No. 2, pp.247-255, April 1967.
- (6) Fralick, S. C. and R. W. Scott, "Nonparametric Bayes-risk Estimation", IEEE Trans. on Information Theory, Vol. IT-17, No. 4, pp.440-444, July 1971.
- (7) Fraser, D. A. S., Nonparametric Methods in Statistics, New York: John Wiley & Sons, Inc., pp.289-292, 1957.
- (8) Fu, K. S., Sequential Methods in Pattern Recognition and Machine Learning, New York: Academic Press, pp.97-101, 107, 1968.
- (9) Fu, K. S. and Y. T. Chien, "Sequential Recognition Using a Non-parametric Ranking Procedure", IEEE Trans. on Information Theory, Vol. IT-13, No. 3, pp.484-492, July 1967.

- (10) Groeneveld, R. A. "A Nonparametric Rank Correlation Method for Detecting Signal in Additive Noise", IEEE Trans. on Information Theory (Correspondence), Vol. IT-13, No. 2, pp.315-316, April 1967.
- (11) Hajek, J., Nonparametric Statistics, New York: Holden-Day, 1969.
- (12) Hajek, J. and Z. Sidak, Theory of Rank Tests, New York: Academic Press, 1967.
- (13) Henrichon, E. G. and K. S. Fu, "Calamity Detection Using Nonparametric Statistics", IEEE Trans. on Systems, Men, and Cybernetics, Vol. SSC-5, No. 2, pp.150-155, April 1969.
- (14) Hoel, P. G., Introduction to Mathematical Statistics, New York: John Wiley & Sons, Inc., pp.329-349, April 1967.
- (15) Hodges, J. L., and E. L. Lehmann, "The Efficiency of Some Nonparametric Competitors of the t-test", Annals of Math. Statistics, Vol. 27, pp.324-335, 1956.
- (16) Kanefsky, M. and J. B. Thomas, "On Adaptive Nonparametric Detection Systems Using Dependent Samples", IEEE Trans. on Information Theory, Vol. IT-11, No. 4, pp.521-526, October 1965.
- (17) Kraft, C. H. and C. van Eeden, A Nonparametric Introduction to Statistics, New York: McMillan, 1968.
- (18) Millard, J. B. and L. Kurz, "Adaptive Threshold Detection of M-ary Signals in Statistically Undefined Noise", IEEE Trans. on Information Theory (Correspondence), Vol. IT-13, No. 2, pp.341-342, April 1967.
- (19) Millard, J. B. and L. Kurz, "Nonparametric Signal Detection - An Application of the K-S, Cramer-von Mises Tests", Report TR

400-127, New York University Laboratory for Electrosence
Research, New York: January 1966.

- (20) Noether, G. E., Elements of Nonparametric Statistics, New York:
John Wiley & Sons, Inc., 1967.
 - (21) Savage, I. R., Bibliography of Nonparametric Statistics,
Harvard University Press, 1962.
 - (22) Thomas, J. B., "Nonparametric Detection", Proceedings of IEEE,
Vol. 58, No. 5, pp.623-631, May 1970.
 - (23) Walsh, J. E., Handbook of Nonparametric Statistics, Princeton,
N. J.: Van Nostrand, pp.158-165, 398-405, 1962.
 - (24) Woinsky, M. N., "Nonparametric Detection Using Spectral Data",
IEEE Trans. on Information Theory, Vol. IT-18, No. 1, pp.110-
118, January 1972.
 - (25) Wolff, S. S., J. B. Thomas and T. R. Williams, "The Polarity
Coincidence Correlator; a Nonparametric Detection Device",
IRE Trans. on Information Theory, Vol. IT-8, January 1962.
- B. On General Probability and Pattern Recognition
- (26) Cooper, P. W., "Quadratic Discriminant Functions in Pattern
Recognition", IEEE Trans. on Information Theory (Correspondence),
Vol. IT-11, No. 2, pp.313-315, April 1965.
 - (27) Cover, T. M. and P. E. Hart, "Nearest Neighbor Pattern Classifi-
cation", IEEE Trans. on Information Theory, Vol. IT-13, No. 1,
pp.21-27, January 1967.
 - (28) Hancock, J. C. and P. A. Wintz, Signal Detection Theory, New York:
McGraw-Hill, pp.209-210, 1966.

- (29) Ho, Y. C. and A. K. Agrawala, "On Pattern Classification Algorithms-Introduction and Survey", Proceedings of IEEE, Vol. 56, December 1968.
- (30) Kashyap, R. L. and C. C. Blaydon, "Recovery of Functions from Noisy Measurements Taken at Randomly Selected Points and Its Application to Pattern Classification", Proceedings of IEEE, 1966.
- (31) Koch, G. S. and R. F. Link, Statistical Analysis of Geological Data, New York: John Wiley & Sons, Inc., 1970.
- (32) Lewis, A. J. and H. C. MacDonald, "Interpretive and Mosaicking Problems of SLAR", Remote Sensing of Environment, Vol. 1, No. 4, pp.231-236, December 1970.
- (33) Nelson, G. D. and D. M. Levy, "A Dynamic Programming Approach to the Selection of Pattern Features", IEEE Trans. on Systems, Science, and Cybernetics, Vol. SSC-4, No. 2, pp.145-147, July 1968.
- (34) Nelson, G. D. and D. M. Levy, "Selection of Pattern Features by Mathematical Programming Algorithms", IEEE Trans. on Systems Science, and Cybernetics, Vol. SSC-6, No. 1, pp.20-25, January, 1970.
- (35) Nilsson, N. J., Learning Machines, New York: McGraw-Hill, 1965.
- (36) Papoulis, A., Probability, Random Variables, and Stochastic Processes, New York: McGraw-Hill, 1965.
- (37) Schwartz, M., Information Transmission, Modulation and Noise, 2nd Ed. New York: McGraw-Hill, pp.366-368, 1970.
- (38) Wee, W. G., "A Survey of Pattern Recognition", T-249, S&RD, Honeywell Inc., St. Paul, Minn.

- (39) Zagalsky, N., "A New Formulation of a Classification Procedure",
M.S. Thesis, University of Minnesota, March 1968.